

A Daily-level Purchasing Model at an E-commerce Site

Shota Sato*, Yumi Asahi**

* Department of Engineering Management Science, Tokyo University of Science Japan

** Management of Business Development, Shizuoka University Japan

Article Info

Article history:

Received Jul 24, 2012

Revised Nov 19, 2012

Accepted Aug 29, 2012

Keyword:

Choice Model

Clickstream Data

Hierarchical Bayes Model

One-To-One Marketing

The Internet

ABSTRACT

Recently, the e-commerce sites market in Japan is growing. Many companies in e-commerce sites compete for customers each other. Therefore, e-commerce managers need to find good customers who may benefit before these customers purchase at rivals. Some preceding studies discuss which visits lead to purchase. However, it is unclear when a customer visits. In our approach, we predict the day a customer will purchase. As a result, we found some customers who tend to purchase on the day after they viewed many product pages. We also found that product page viewing before purchasing may be one of good predictors.

Copyright © 2012 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Shota Sato,

Department of Engineering Management Science,

Tokyo University of Science

Email: s_sato@ms.kagu.tus.ac.jp

1. INTRODUCTION

In Japan, a penetration of the Internet is 80.0% in 2011[6]. Most people in Japan use the Internet to investigate something, send an e-mail, purchase something and so on. Recently, the e-commerce sites market in Japan is growing [8]. Many companies will compete in the e-commerce market. Many companies in e-commerce market compete for customers with each other. Therefore, e-commerce managers need to decide which customer is good customer who may benefit in the near future in order to win many rivals. In other words, e-commerce managers need to understand customer's purchasing behavior at e-commerce sites, and they need to predict which customer may benefit.

Customer's purchasing in e-commerce sites is one of the major problems for e-commerce managers. Several preceding studies focused on customer's purchasing in e-commerce sites. For example, Moe and Fader [10] developed a model which predicts customer's purchasing. They predict customer's probability of purchasing based on an observed history of visits and purchases. Bucklin and Sismeiro[5] proposed a model which predicts customer's purchasing. They decomposed the user's purchase process into the several steps such as "completion of product configuration", "input of complete personal information". Van den Poel and Buckinx [12] analyzed purchasing behavior and they found that variables such as "Number of days since last visit" have strong effect on purchasing. These studies discuss which visit leads to purchase. However, it is unclear when a customer visits. Predicting the day a customer visits and purchases helps e-commerce managers to understand when a customer intends to purchase.

To predict customer's purchase, it is important for e-commerce managers to understand consumer behavior which leads to purchase. Moe and Fader [10] focus on a history of visits and purchases to predict purchasing. We focus on viewing of product pages as behavior which leads to purchase. As a preceding study which focus on product pages, Moe [9] found direct-buying visit exhibits more focused behavior with more page views devoted to product-level information. However, the relationship between purchasing behavior

and viewing of product pages before purchasing were not revealed. Some customers may gather information about products for several days before purchasing. The other customers may make blind purchase without viewing many products pages. Customers who begin to gather information about products several days before purchasing can be a good target because it is easy for managers to predict their purchasing according to their behavior. That is, finding customers who begin to gather information some days before purchasing can be important matter for managers.

It is speculated that behavior which leads to purchase is different for each customer. Some customers may purchase after gathering much information. Other customers may purchase long time after their last purchase. Capturing individual-level behavior helps managers to get insights for one-to-one marketing. Therefore we aim to reveal customer's individual-level behavior which leads to purchase.

E-commerce managers need to win many rivals. Customer's purchasing in e-commerce sites is one of the major problems for e-commerce managers. Preceding studies [5][10][12] predict which visit leads to purchase. However, it seems unclear when a customer visits. In order to solve this problem, we aim to predict the day customer will visit and purchase. For example, we predict whether a customer will purchase or not tomorrow. We also aim to capture customer's heterogeneity. Purposes of this study are to develop a model which predicts customer's purchasing behavior and to reveal the behavior which leads to purchase. To get managerial insights on marketing strategies to customers is also the purpose of this study.

2. RESEARCH METHOD

In this study, we develop the model which focus on customer's purchasing. A binary variable y_{it} equals one if customer i ($i=1, \dots, N$) purchase on the t th day ($t=1, \dots, T$) and zero the customer didn't purchase. We employ binary probit model. Let u_{it} be the utility of customer i on t th day. A binary variable y_{it} equals one if u_{it} exceeds zero otherwise y_{it} equals zero. The utility u_{it} is specified as

$$u_{it} = x'_{it} \beta_i + \varepsilon_{it} \quad \varepsilon_{it} \sim N(0,1) \quad (1)$$

where x_{it} is a vector of independent variables which contain an intercept. β_i is a parameter vector of customer i . In this study, a parameter vector β_i is different for each customer in order to capture heterogeneity.

To capture heterogeneity, hierarchical Bayes model is employed. Parameter vector β_i is specified as

$$\beta_i = \Delta + \eta_i \quad \eta_i \sim MVN(0, V_\beta) \quad (2)$$

where Δ is parameter, and $MVN(0, V_\beta)$ is multivariate normal distribution with a zero mean vector, covariance matrix V_β . Multivariate normal distribution is used as a prior distribution of Δ . Inverse Wishart distribution is used as a prior distribution of V_β . Multivariate normal prior and inverse Wishart prior are written as

$$\Delta \sim MVN\left(\bar{\Delta}, A\right) \quad (3)$$

$$V_\beta \sim IW(v_0, V_0) \quad (4)$$

where IW is inverse Wishart distribution. Constants $(\bar{\Delta}, A, v_0, V_0)$ are chosen to provide a diffuse prior for parameters Δ and V_β [11].

3. DATA

Data of Joint Association Study Group of Management Science was used. This data contains clickstream data, transaction data of an e-commerce site which sells golf products in Japan. Data was collected from July 1, 2010 to June 28, 2011. Figure 1 shows data summary of the total sale at weekly-level. Figure 1 was drawn as 1 dollar equals 80 yen because one dollar equals around 80 yen on July 1st. The vertical axis in figure 1 indicates the total sale in data and the horizontal axis indicates week. Figure 1 shows the total money decreases at the 37th week. The 37th week is from March 10, 2011 to March 17, 2011. In Japan, a massive earthquake occurred in March 11, 2011. Because of a massive earthquake, customers may abstain from playing golf and purchasing golf products. Therefore Data from July 1, 2010 to February 28, 2011 were used. We use data from July 1, 2010 to January 31, 2011 as an estimation period and the remaining period was set aside as a holdout period. Customers are qualified for inclusion in the sample if they purchase something more than on 4 days in estimation period. The number of customers we analyzed is 157. In this study, we summarized data at daily-level because we aim to capture customer's purchasing behavior each day.

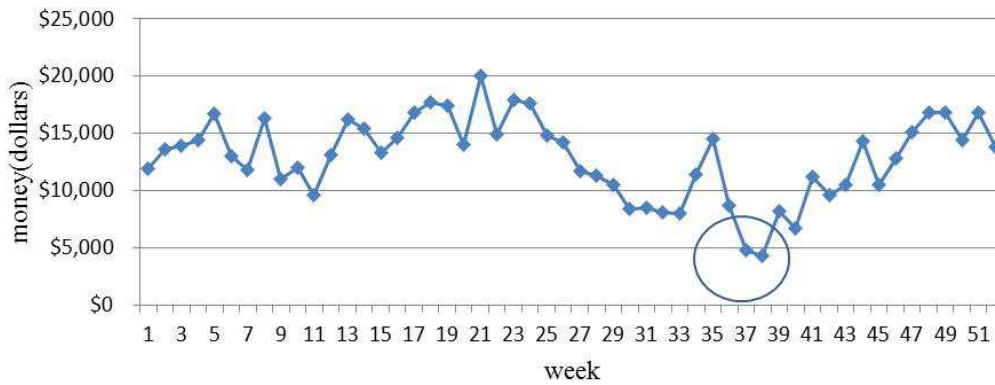


Figure 1. The weekly-level total money

A rate of male in data, an average of age and an average of the number of days a customer purchased are shown in table 1 as a summary of data. Table 1 shows most customers in data we used are male. Data we used in this study is data of an e-commerce company which sells golf products. This may be a reason most customers are male. An average of customer’s age in data we used is around 50. An average of the total number of days a customer purchased is around 7.

Table 1. A summary of data

a rate of male	average of age	Average of the number of days a customer purchased
93.60 %	49.15	7.00

In this study, we use some variables as independent variables x_{it} . Variables we use in this study are shown in table 2. The variable CumPages is contained in the model to capture the relationship between purchasing behavior and the total number of product pages which a customer viewed in the past 7 days. We believe the variable CumPages has positive effect on purchasing. Moe and Fader[10] predict customer’s purchase based on an observed history of visits and purchases. Therefore, the variable, CumVisit, is contained in the model as the history of site visit. CumVisit is contained in the model to capture the relationship between purchasing behavior and the total number which a customer visit the site in the past 7 days. The variable CumMoney is contained in the model to capture the relationship between purchasing behavior and total of money which a customer spent in the past 30 days. We believe the variable CumMoney has negative effect on purchasing. For example, customers who spent much money in past several weeks may abstain from purchasing for a while.

CumMoney was defined as the 10^{-4} yen. Logarithmic of CumPages after 1 was added and logarithmic of CumVisit after 1 was added were used because logarithmic of page view and logarithmic of repeat visit were used in a preceding study[4].

Table 2. The details of independent variables

Name Of Variable	Description
CumPages	the total number of product pages which customer i viewed between the (t-7) th day and (t-1) th day
CumVisit	the total number which customer i visit the site between the (t-7) th day and (t-1) th day
CumMoney	the total of money which customer i spent between the (t-30)th day and (t-1) th day

4. RESULTS AND ANALYSIS

4.1. Estimation method

Parameters are estimated using Markov Chain Monte Carlo (MCMC) method. A Gibbs sampler is used to draw samples of parameters β_i, Δ, V_β . We use 100,000 draws for burn in, and additional 100,000 draws were used to infer the posterior distribution of the parameters. We kept 5 th draw to reduce computer memory requirement. The resulting 20,000 draws were used in our analysis. Posterior means are used as estimates.

4.2. Model camprison

First, the performance of models is investigated and some models are compared. The performance of models is investigated by Receiver Operation Characteristics (ROC) curve and Area under the curve (AUC)[2]. ROC curve is drawn by Table 3.

Table 3. A matrix of prediction and result

Prediction\Result	Purchase	Not Purchase
purchase	True Positive(a)	False Positive(b)
not purchase	False Negative(c)	True Negative(d)

ROC curve is a figure which False Positive Rate ($b/(b+d)$) and True Positive Rate ($a/(a+c)$) are plotted as the threshold is changed. Area under the ROC curve is AUC. AUC is 1 at the maximum and that is 0.5 in case of predicting randomly. AUC indicates that the closer to 1 AUC of the model is, the better performance the model is. Purchasing possibility is calculated by estimates and data of holdout period. Therefore, if AUC of the model exceed 0.5, it is interpreted that the performance of the model is better than random prediction. We calculate purchase possibility of customer i on t th day, p_{it} , as

$$p_{it} = \Phi(x_{it}\beta_i) \quad (5)$$

where $\Phi()$ is cumulative density function of standard normal distribution. We compare the model of this study with the logit model without heterogeneity used in a study of Van den Poel and Buckinx[12]. The logit model without heterogeneity is written as

$$p_{it} = \frac{\exp(u_{it})}{1 + \exp(u_{it})} \quad (6)$$

$$u_{it} = x_{it}\beta_i \quad (7)$$

where p_{it} is the probability of purchase of customer i on the t th day. Independent variables x_{it} contains same variables as our model. We estimate parameters of logit model without heterogeneity by maximum likelihood method.

We compared some models. Model 1 is hierarchical Bayes binary probit model with a variable of product pages. Model 1 is our approach. Model 2 is hierarchical Bayes binary probit model with a variable of site visits. Moe and Fader[10] predict customer's purchase based on an observed history of visits and purchases. Therefore, Model 2 contained the variable of site visits instead of product page viewing. Model 3 and Model 4 are approaches of Van den Poel and Buckinx[12]. They predict the possibility of purchase on customer's next visit using traditional logit model without heterogeneity, they don't predict the day which a customer will visit and purchase. Their approach is easily to be applied to the prediction of the day of visit and purchase because they employ the traditional logistic regression. Therefore we compare our approach and their approach. Our approach aim to capture customer's individual heterogeneity though an approach of Van den Poel and Buckinx doesn't capture customer's individual heterogeneity. The performance of models is shown in Table 4.

Table 4. The performance of models

	variables			AUC	
	CumPages	CumVisit	CumMoney	estimate period	holdout period
Model 1	✓		✓	0.723	0.689
Model 2		✓	✓	0.716	0.683
Model 3 (Preceeding study's approach[12])	✓		✓	0.621	0.665
Model 4 (Preceeding study's approach[12])		✓	✓	0.607	0.655

The model which has larger AUC has the better performance. The performance of Model 1, our approach with a variable of product pages, of estimate period was 0.722 of AUC and that of holdout period was 0.689 of AUC. The performance of our approach with a variable of site visits, Model 2, of estimate period was 0.716 of AUC and that of holdout period was 0.683 of AUC. This result shows that Model 1 has better performance than Model 2 because AUC of Model 1 is larger than that of Model 2. This result implies that product page viewing may be more useful for predicting customer's purchase than a history of visits.

The performance of Model 3, preceding study's approach[12], of estimate period was 0.621 and that of holdout period was 0.665. The performance of Model 4, preceding study's approach[12], of estimate period was 0.607 and that of holdout period was 0.655. Especially, Model 3 and Model 4 has not good performance in estimate period.

The result of AUC shows our approach is better than random prediction because AUC of that exceeds 0.5. AUC of our approach is larger than that of preceding study's approach[12]. Therefore, performance of our model is better than the traditional model. In this study, we discuss Model 1 because Model 1 has the best performance in 4 models.

4.3. Result

First, table 5 shows posterior means of parameters Δ . Parameters Δ indicate the average effect on purchase. The sign "***" in table 5 indicates the 95% confidence interval of the parameter doesn't contain zero. Table 5 shows the variable CumPages has positive effect on purchasing on average. This result indicates the number of product pages which a customer viewed in the past several days has positive effect on purchasing. This result suggests that a customer view many product pages several days before purchasing. The variable CumMoney has negative effect on purchasing. This means a customer tend not to purchase after he/she spent much money. A customer will not purchase several weeks after he/she spent much money because he/she may hesitate to spend money after spending much money.

Table 5. Esitemates of

Intercept	CumPages	CumMoney
-2.13**	0.15**	-0.16**

Next, individual-level parameters are analyzed. Our model estimates individual-level parameters. Figure 2 shows a histogram of each customer's parameter of CumPages. Figure 2 shows that parameters of CumPages of most customers are around 0.15. This result suggests that the total number of product pages which a customer viewed in the past several days has slight positive effect on purchasing for most customers.

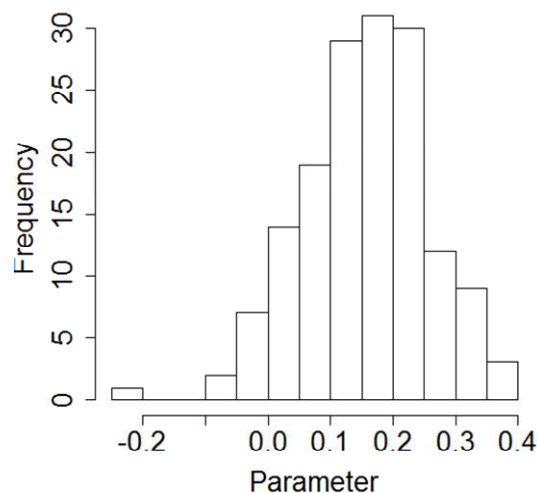


Figure 2. A histogram of CumPages

Figure 2 also shows the variable CumPages has positive effect on purchasing for some customers. It is interpreted that customers whose parameter of CumPages is positive tend to gather information about products several days before purchasing something. It will be easy to predict purchasing behavior of these customers because they tend to purchase several days after viewing many product pages. The variable CumPages has no or negative effect on purchasing for some customers. It is speculated that these customers tend to purchase on the day they viewed many product pages or tend to make blind purchase.

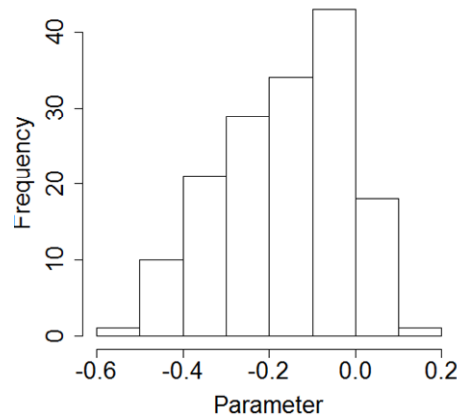


Figure 3. A histogram of CumMoney

Figure 3 shows a histogram of each customer's parameter of CumMoeny. Figure 3 shows that parameter of CumMoney of most customers is around -0.2. This result suggests that the total of money which a customer spent in the past 30days has negative effect on purchasing for most customers. Figure 3 also shows parameters of CumMoney of some customers are around zero or positive. It is interpreted that customers whose parameter of CumMoney is positive tend to purchase several weeks after they purchase something. It is speculated that these customers tend to continue to have an interest in golf in a short period because they tend to purchase several weeks after purchasing.

The number of customers whose 95% and 90% confidence intervals of parameters don't contain zero is shown in table 6. The total number of customers analyzed in this study is 157. 30 customers have positive parameter of CumPages and their 95% confidence intervals of the parameter don't contain zero. Table 6 shows that around one-fifth customers in this data have strong relationship between purchasing and viewing product pages. The 90% confidence intervals of 42 customers don't contain zero. That is, they have a tendency to purchase several days after viewing many product pages. 14 customers have negative parameter of CumMoney and the 95 % confidence intervals of the parameter don't contain zero. The 90% confidence intervals of 26 customers don't contain zero. That is, they have a tendency not to purchase several weeks after spending much money.

Variables	Confidence interval	
	95%	90%
CumPages	30	42
CumMoney	14	26

Next, the relationship between CumPages and total money which a customer spent in data period is analyzed. The reason we focus on CumPages is that a parameter of CumPages is useful for predicting purchasing because customers whose parameter of CumPages is high tend to purchase several days after viewing many product pages. A customer who spent much money in the e-commerce site is a good customer for the e-commerce company and e-commerce managers will promote them to purchase and aim customer retention. However, some good customers who spend much money may tend to purchase several days after viewing product pages. Other good customers may tend to make blind purchase. In other words, good

customer's behavior which leads to purchasing at the e-commerce site may be different. Understanding customer behavior before purchasing such as viewing product pages will provide managerial insights.

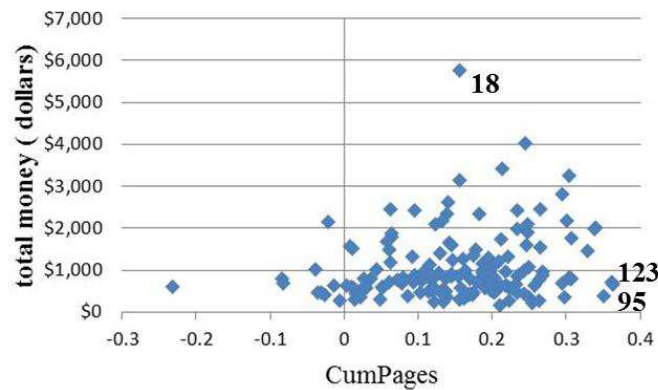


Figure 4. A scatter plot of total money and CumPages

Figure 4 shows a scatter plot of total money a customer spent in data period (total money) and the parameter of CumPages. Dots in figure 4 mean customers. The number attached to the dot which means each customer indicates the serial number of customer. The vertical axis in figure 4 means total money (dollars) a customer spent in data period. Figure 4 was drawn as 1 dollar equals 80 yen. Customers who are located in the upper spent more money in data period. The horizontal axis in figure 4 means the parameter of CumPages. Customers located in the left have the higher parameter of CumPages. Therefore, a customer plotted in the more right in figure 4 is more likely to purchase after viewing many product pages.

A few characteristic customers are picked up as examples. Customer No. 18 is located in the upper middle in figure 4. Customer No. 18 is the best customer in respect of total money. Customer No. 123 and customer No.95 are located in the lower left in figure 4. They spent less money in data period. They have a strong tendency to purchase something after viewing many product pages. Therefore they can be one of the best targets in respect of predicting purchase according to viewing product pages and can be a passable good target in respect of total money.

Analysis using the model of this study considering heterogeneity reveals the individual-level tendency before purchasing. The traditional logit model used in a preceding study[12] without heterogeneity reveals the whole tendency which leads to purchase but doesn't reveal the individual-level tendency. The individual-level tendency provides managerial insights for one-to-one marketing. By revealing the individual-level tendency before purchasing, e-commerce managers can get insights about which customer will be a good target and when a customer intends to purchase.

5. CONCLUSION

Customer's purchasing in e-commerce sites is one of the major problems for e-commerce managers. Preceding studies discuss which visit lead to purchase to solve this problem. However a customer's next visit may occur 5 minutes after or may occur 1 month after. In other words, it is unclear when a customer visits. In order to solve this problems, in this research, we develop a model which predicts the day a customer will purchase considering customer's individual-level heterogeneity.

There are four findings as results of this research. First, a model which predicts the day a customer will purchase was developed and our approach has a better performance than the traditional model. Second, we found that a model which contains product page viewing and spending money as predictors has better performance than a model which contains a history of visit and spending money as predictors. Third, we found that viewing product pages has positive effects on purchase on average and spending money past a month has negative effects on average. Last, Some customers who tend to purchase after viewing many product pages were found.

First, we developed hierarchical Bayes binary probit model considering heterogeneity. We compared the performance of the model of this study with the performance of preceding study's approach[12] and that of random prediction. As a result of comparing the performance of the model by using AUC, the performance of the model of this study is better than preceding study's approach[12] and random prediction in both the estimation period and the holdout period.

A preceding study[10] predicted customer's purchase based on a history of visits and purchases. In this study, we predict customer's purchase by product page viewing instead of a history of site visits. As a result, a model which contains product page viewing and spending money has better performance than a model which contains a history of visit and spending money. This result implies that product page viewing may be a useful predictor for purchase in several days after.

Next, the whole effect on purchasing was investigated. It is found that the variable CumPages, the total number of product pages a customer viewed in the past 7 days, has a positive effect on purchasing on average. Customers more or less tend to view product pages and gather information before purchasing. We also found that the variable CumMoney, the total of money a customer spent in the past 30 days, has a negative effect on purchasing on average. Customers will not purchase something after they spent much money. In this research, we found behavior which leads to purchase at an e-commerce site.

Last, the individual-level parameters were investigated. By investigating the individual-level parameters, some customer's behavior which leads to purchase is revealed. As a result of analysis, we found some customers tend to purchase several days after viewing many product pages. It is found that around one-fifth customers in data we analyzed have strong relationship between purchasing and viewing product pages before purchase. These customers can be good targets in respect of predicting purchasing according to their viewing behavior. We also found some customers tend not to purchase several weeks after they spent much money. We investigated which customer can be a good target by a scatter plot of total money and the parameter of a predictor of CumPages. Preceding study's approach doesn't reveal individual-level tendency which leads to purchase. In this research, we found individual-level tendency which leads to purchase.

As future studies, the accuracy of prediction of purchase may need to be improved by the variables about the frequency of playing golf. Data in this study is clickstream data and transaction data of an e-commerce site which sells golf products. Therefore, the variable like "the total number of days since a customer played golf last" may contribute the prediction of purchasing. Empirical analysis on other e-commerce sites such as a site which sells clothes will be needed and the result need to be compared with the result of this study. In this study, we developed the model which capture customer's purchasing behavior on daily level. Purchasing of a customer is never made in the e-commerce site unless a customer visits the site. Capturing both visiting and purchasing behavior will provide useful knowledge. That is, to enhance the model to capture customer's visiting behavior and purchasing behavior will be a future task. Some studies[1][3] developed the model which integrate customer's purchase incident, brand choice and purchase quantity. To apply their model to clickstream data, to capture tendencies of visiting and purchasing behavior will be a future task.

Clickstream data and transaction data in the e-commerce site were used in this study. However, some e-commerce companies have not only a virtual shop but also a real shop. In this study, data of a virtual shop was only analyzed. Kimery and Amirkhalkhali[7] found that trustworthiness is strong motivators for e-commerce adoption for Japanese. Therefore some Japanese customer may like to purchase at a virtual shop, but others may like to purchase at a real shop. Some Japanese may feel trustworthy to purchase at a real shop. Therefore, integrating data of virtual shop and data of real shop, analyzing them will provide more useful insights for managers to understand customer's behavior. Analysis using integrated data which contains both data of virtual shop and data of real shop will also be a future task.

ACKNOWLEDGEMENTS

We would like to thank to professor Yamaguchi, T. (Department of Management Science, Faculty of Engineering, Tokyo University of Science) for his support on our experiment.

REFERENCES

- [1] Andrews. R. L, and Currim. I. S, "Multi-stage purchase models: Accommodating response heterogeneity, common demand shocks and endogeneity using disaggregate data", *International Journal of Research in Marketing*, vol. 26, pp. 197-206, 2009.
- [2] Blattberg. R. C, Kim, B and Neslin. S. A, "Database Marketing: Analyzing and Managing Customers", Springer, 2008.
- [3] Bucklin. R. E. and Gupta. S, "Brand choice, purchase incidence, and segmentation: An integrated modeling approach", *Journal of Marketing Research*, Vol. 29, pp. 201-215, 1992.
- [4] Bucklin. R. E. Sismeiro. C, "A Model of Web Site Browsing Behavior Estimated on Clickstream Data", *Journal of Marketing Research*, Vol. 40, No. 3, pp. 249-267, 2003.
- [5] Bucklin. R. E. Sismeiro. C, "Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach", *Journal of Marketing Research*, Vol.41, No. 3, pp. 306-323, 2004.
- [6] Internet World Stats, Retrieved August 13, 2012, from the world wide web: <http://www.internetworldstats.com/>

- [7] Kimery. M. K and Amirkhalkahali. S, “Cultural Influences on The Adoption Of Electronic Commerce: Initial Results From Japan And The United States”, *International Business & Economics Research Journal*, Vol. 8, No. 6, pp. 41-51, 2009.
- [8] Ministry of Economy, Trade and Industry of Japan, Retrieved December 11th , 2012 from the World Wide Web: http://www.meti.go.jp/policy/it_policy/statistics/outlook/bessi3H23EChoukokusho.pdf.
- [9] Moe. W. W, “Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstreams”, *Journal of Consumer Psychology*, 13,pp. 29-39, 2003
- [10] Moe. W. W, Fader. P. S “Dynamic Conversion Behavior at E-Commerce Sites” , *Management Science* ,Vol. 50, No. 3, pp.326-335, 2004
- [11] Rossi. P. E, Allenby. G and McCulloch. T, “Bayesian Statistics and Marketing” , John Wiley & Sons, Jersey, 2005.
- [12] Van den Poel. D, Buckinx. W, “Predicting online-purchasing behavior” , *European Journal of Operational Research*, Vol. 166, No. 2, pp. 557-575, 2005.

BIOGRAPHIES OF AUTHORS

Shota Sato received a BEng and MEng in management science from Tokyo University of Science, Japan in 2011 and 2013, respectively. His area of research is focusing on marketing science.



Yumi Asahi belongs to the Faculty of Engineering of Shizuoka University. She is an Associate professor at Management of business development. She has got Ph.D. of Engineering from Tokyo University of Science. Her area of research is focusing on behavior metrics and marketing.