Analysis of Impacts on the Purchasing Behavior Using Information on Social Networking Services

Kohei Otake†
Faculty of Science and Engineering
Chuo University, Tokyo, Japan
Tel: (+81) 3-3817-1939, Email: otake@indsys.chuo-u.ac.jp

Takashi Namatame
Faculty of Science and Engineering
Chuo University, Tokyo, Japan
Tel: (+81) 3-3817-1932, Email: nama@indsys.chuo-u.ac.jp

Abstract. Recently, Social Networking Services (SNSs) have been growing in popularity. There have been many studies about marketing activities (e.g. consumer behavior or sales promotion) using SNSs. In this study, we aim to clarify the relationship between SNSs and purchasing on the EC (electronic commerce) site. Especially, in this paper we indicate the results of predictive models of purchasing behavior that was created using the information on Twitter. We focused on Twitter that is one of the most popular SNSs. Moreover, we focused on the golf portal site that is the subject of this research. Firstly, we conducted a correlation analysis with respect to purchasing history and tweets. As the results, it became clear that "There is a weak positive correlation between the number of purchase and the number of tweets in 1 year" and "Tweets have strong positive correlation with a specific month, products and tweets." Secondly, through the results of correlation analysis, we create 768 predictive models of purchasing behavior by logistic regression. Finally, we conducted hierarchical cluster analysis using each the parameters of each model. We analyzed the differences of these clusters. It was found that users have distinctive attribute information on Twitter were improved number of purchase.

Keywords: Social Networking Service, Marketing, Purchasing Behavior, Logistic Regression

1. INTRODUCTION

Recently, social media have been growing in popularity. Social media is the collective of online communications channels with community-based input, content-sharing, interaction and collaboration with the other user. Websites and applications dedicated to forums, social networking, social bookmarking and wikis are among the different types of social media. Especially, the users of Social Networking Services (SNSs) are increasing among the social media. Users can communicate with friends about their preference and favorites using SNSs. Under such situation, marketing campaigns using SNSs have been receiving increasing attention in EC (Electronic Commerce) suppliers. SNSs have variety of information of products such as product introductions, utilization ways and reviews. Additionally, this information has a few features such as "spread area is wide" and "communication speed is fast". Therefore, using information of SNSs effectively is an important subject in EC suppliers.

There have been many studies about marketing activities (e.g. consumer behavior, sales promotion) using SNSs in this situation. However, prior attempts to measure the effects of SNSs have been inconclusive. Moreover, these previous research did not consider the background of users who post the information in SNSs. There are accounts using SNSs as communication tools with their friends. On the other hand there are accounts (e.g. company's account, well-known person about the contents topic) using SNSs as advertising tool about products, brands. Influence is different even if both of them send the same information. We think that influence to purchase of products is different from the account's background information (e.g. celebrity, belonging).

2. PURPOSE OF THIS STUDY

In this study, we aim to clarify the relationship between activity on SNSs and purchasing on the EC (electronic commerce) site. Especially, in this paper we indicate the results of predictive models of purchasing behavior that was created using the information of account's background. We focused on Twitter that is one of the most popular SNSs.

We focused on the golf portal site that is the subject of this research. We were offered marketing data such as consumer information, access log of EC site and purchasing information by the golf portal site. Moreover, we collected tweets about golf activities from Twitter. We call these tweets "Social data". In this study, we investigate to create appropriate models of purchasing behavior by using marketing data and social data.

3. PREVIOUS RESEARCH

In previous research, consumer's direct behavior data such as access history and POS (Point of sales) data has been often used to create consumer behavior models (Gupta, 1988, Shaw, et al., 2001) Some such examples include a study (Moe and Fader, 2004) proposed purchase occurrence probability model, a study (Koike, et al. 2007) that tried discovery of important customers, and a study (Hisamatsu, 2013) that created models which predict at the timing of purchase.

On the other hand, a few studies included Social media information to purchasing history and access history. These representative studies addressing behavior on social media and purchase behavior concern movie box-office records. Some such examples include a study (Liu, 2006) that divided the content of comments posted to YAHOO! Movies about a certain movie into positive and negative and analyzed the connection with that movie's box-office record, a study Mishne nad Glance, 2006) analyzing the effect of negative posts on social blogs on a movie's box-office record, and a study (Yoshida, et al., 2007) that modeled the combined effect on movie box-office records of both social blog posts and the volume of television advertisement. As Tsurumi et al. (2013) have pointed out, it is thought that the reason these studies focus on movie box-office records is the ease of access to such box-office records and the related text data of reviews and comments posted by consumers.

From the results of previous research, we focused on following two points in this study.

1. Based on the idea that social data indicates protuberance in a market, we regard social data as "Trend Variable."
2. We focus on all products on the EC site to the subjects of this research. Moreover, we set analysis period to one year.

We catch SNSs with "Trend Variable" like as Tsurumi et al. (2013, 2015). We think that it is possible to create appropriate models of purchasing behavior by adding "Trend Variable" from SNSs.

Previous researchers analyzed only a few limited products and periods about 2. However, when considering marketing campaigns of the whole EC site, it is necessary to analyze products of wide range and various categories. In this study, we analyzed 64 product categories among 1. We will begin with an explanation of the data set utilized in this study.

4. PROPOSAL OF PURCHASING BEHAVIOR MODEL USING TWITTER INFORMATION

4.1 Data Sets

We will begin with an explanation of the data set utiliz ed in this study.

4.1.1 Marketing Data

We were offered following marketing data by the golf portal site.

- Purchasing History Data
  - Data about purchase of customers who went by EC site (e.g. Product Name, Purchasing Time, Product Price and Category of Product)
- Access History Data
  - Web access data of customers who visited the EC site (e.g. Access time, Access Product, Device and Referrer)

In this study, purchasing history data and access history data from the 12 month interval from June 2012 to May 2013 was used. We defined these data as marketing data. Further we focused on customers who have unique ID.

The outline of marketing data is shown Table.1.

Table 1: Outline of marketing data in this study.

Data Period	2012-06-01~ 2013-05-31
Total number of Purchasing	993,768
Average of Purchasing number per 1 day	2,723
Standard deviation of purchase number per 1 day	550

Table 2: Outline of social data in this study.

ID Attribution		Total Number of Accounts	Total Number of Tweets (in the data period)
All accounts		33	13,094
Person	Golfer	17	7,347
Company	Golf News	5	2,973
	Golf Shop	4	522
	Golf Course	3	2,565
	Golf Sale of Products	1	2,053
	Golf Manufacturer	2	463
	Training	1	1,863

4.1.2 Social Data

We selected 33 accounts based on follow-follower relationship information on the accounts registered with the portal site. These accounts have the different background (such as Golfer, Golf News, Golf Shop, Golf Course, Golf Sale of Products, Golf Manufacturer and Golf Training). In this study, we defined these tweet data as social data. The outline of social data is shown Table 2.

4.2 Preliminary Consideration about Models of Purchasing Behavior

Firstly, we conducted a correlation analysis with respect to marketing data and social data. We performed correlation analysis using the number of purchasing of product categories and the number of tweets of golf accounts.

As the result, following tendencies became clear.

- There is relative weak positive correlation between the number of purchase and the number of tweets in 1 year.
- Tweets have strong positive correlation with a specific month, products and tweets.

Through these results, we created the variables for models of purchasing behavior in 1 month and product category. We create purchasing behavior models at the next section.

4.3 Model Making of Purchasing Behavior

In this section, we describe our purchasing behavior m odels using marketing data and social data. The regression model is shown below as formula (1).

$$p_{ijlt} = \frac{1}{1+\exp\{-(\beta_{0lt}+\sum_{c=1}^{k}\beta_{clt}x_{cijlt})\}} \quad (1)$$

Where, p_{ijlt} is the probability that l of product is purchased in site visit day j in customer i to period t. Moreover, β_{0lt} is the intercept and β_{clt} is the coefficient of each explanatory variable.

We set the presence of customer's purchase data as objective variable every visit to site. Additionally, we set marketing data and social data as explanatory variables. Marketing data include 27 explanatory variables such as "Device Name", "Domain", "Way of Access", "Inflow Process", "Visit Times", "Number of Migration", "Number of PV", "Stay Time", "Last Purchase Days" and "Last Session Days". Social data include 33 explanatory variables that collected tweets (Table 2). The models of purchasing behavior were made every 1 month (12 months) and product categories (64 product categories). So, we made 768 models in this study.

Furthermore, we need to compare the value of explanatory variables in the models with social data and marketing data. Therefore, we used standardized explanatory variables. Our making models have 60 explanatory variables. We checked multicollinearity using VIF value (threshold is 10). Moreover, we went stepwise selection using AIC (Akaike's Information Criterion). Glm function of R is used to solve.

Next, in order to verify the effectiveness of models, we performed likelihood ratio test. As the result, 358 models among 768 models were statistically significant. Moreover, we performed 10-fold cross validation to these models. We used AUC (Area Under the Curve) value to decide the optimal parameters. Next section, we focused on 86 models that meet the criteria (AUC more than 0.80 and standard deviation of AUC is less than 0.01).

5. CONSIDERATION OF THE RESULT CONSUMER PURCHASING BEHAVIOR MODELS

First we describe the tendency of 86 models using partial regression coefficients. When we checked the significant partial regression coefficient, individual models

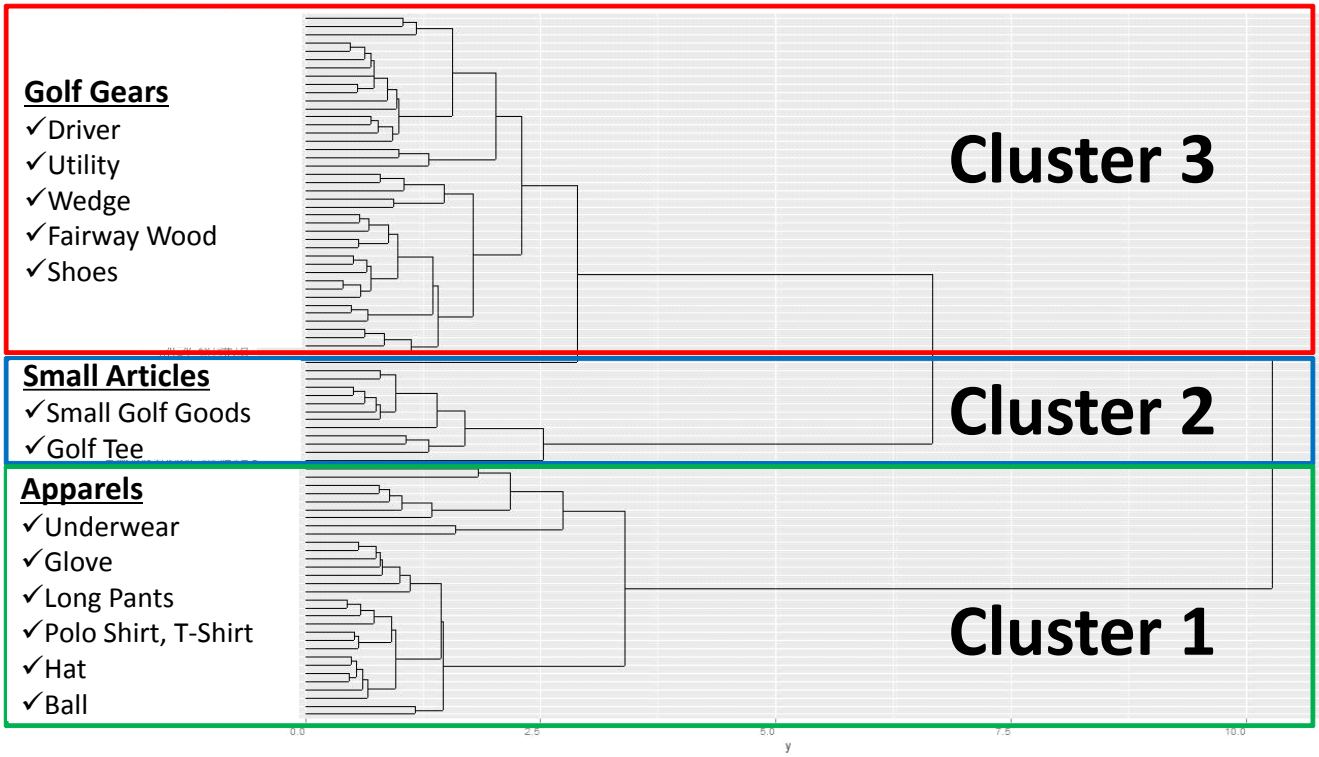

Figure 1: Result of Cluster Analysis.

had 33.66significant explanatory variables in average. Within the significant partial regression coefficient, models had 16.01 positive explanatory variables in average. There were 8.79 variables of marketing data and 7.22 variables of social data as the details of positive explanatory variables. Further, we focus on the mean of significant partial regression coefficient of 86 models. The high values of mean of partial regression coefficient were "Number of migration (mean 1.06)", "Stay Time (mean 0.73)" and "Domain (shop page of site)". We found marketing data had high explanatory force within these models as overall tendency. Moreover social data did not have high explanatory force within these models as overall tendency. However social data had very high explanatory force by several models. For example, professional golfer A had high partial regression coefficient (0.86) in "Polo Shirts/2012" model. Moreover, we checked that golfer A had positive value about other apparel products. Golfer A is a young female golfer and the A's fashion sense is famous for golf funs. We checked A's tweets, we found that A's tweet contained many photo about coordinate of golf fashion. Therefore we think that tweets of A had the positive influence on a purchase probability of apparel products.

Next, we performed cluster analysis to consider the detailed tendency of 86 models. We used partial regression coefficient values for cluster analysis. Further, we adopted hierarchical cluster analysis to make cluster. We used explanatory variables of coefficient of significant probability less than 0.05 to cluster analysis. We used Ward method to make cluster. We set the number of clusters as 3 and consider (Figure 1).

Next, we focused only product categories out from models that belong to the respective clusters and considered each of them. Further, we classified along product categories into large groups of product class. The result is shown in the Table 3. The values are number of models and the percentages of models in each cluster are in the parenthesis. Features obtained from Table 3 shown below.

- ✓ Apparels and Balls kinds of Long-Pants, Polo Shirts and Sweater are more than 80 percent in cluster1. Moreover, all models of Balls are in cluster 1.
- ✓ Small articles kinds of Golf Tee, Ball Case are more than 90 percent in cluster 2.
- ✓ Golf gears kinds of Driver, Putter and Wood are more than 90 percent in cluster 3.

From the result of cluster analysis, it was shown that the respective clusters were formed by the models about particular class of products.

Next, we checked the difference of partial regression coefficient of each clusters. We conducted ANOVA (analysis of variance) using all explanatory variables. From the result, 17 explanatory variables (14 explanatory variables of marketing data and 3 explanatory of variables social data) have unequal distribution for a significance probability of 0.05. The explanatory variables with the suffer difference are shown in Figure 2. The explanatory variables of marketing data had much difference as the tendency of the whole. We consider about the difference of explanatory variables.

Table 3: Result of cluster vs. category

Cluster No / Purchase Types	Cluster1	Cluster2	Cluster3
Iron	1	0	0
Outer	1	0	1
Underwear	1	0	6
Discount Wear Set	1	0	0
Wedge	2	0	1
Caddie Bag	1	0	1
Glove	0	0	8
Shoes	7	0	1
Tee	0	4	0
Driver	7	0	0
Putter	1	0	0
Fairway Wood	3	0	0
Head Cover	0	0	1
Ball	0	0	9
Utility	3	0	0
Small Golf Goods	0	6	1
Long Pants	0	0	5
Practice Goods	2	0	0
Polo Shirt, T-Shirt	1	2	6
Hat	0	0	3
Total Purchases	31	12	43

It became clear that the number of migration is high and the stay time is short about cluster 2. Cluster 2 was formed by small articles models kinds of Golf Tee, Ball Case. Generally small articles are replaced by the high frequency. Moreover volume of information of product page about small articles is little. The models of cluster 2 reflected this tendency appropriately as the tendency of the whole.

On the other hand, it became clear that the number of migration is low and the stay time is long about cluster 3. Cluster3 was formed by Golf gears kinds of Driver, Putter and Wood. Generally golf gears are not replaced by the high frequency. This is because the unit price of golf gear is high. Moreover product page of golf gear have a lot of information about detail of gear (e.g. feeling of a golf club, target ski

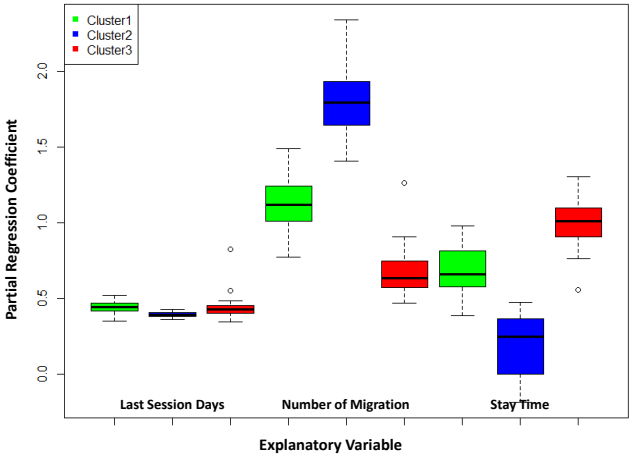

Figure 2: Comparison of some variables among clusters

ll of level). The models of cluster2 reflected this tendency appropriately as the tendency of the whole.

Further, we focused each month out from models that belong to the respective clusters and summarized. However, the particular tendency (e.g. the season) was not found. Therefore, it became clear that partial regression coefficient often shows the feature of the product category.

From the result, we think that partial regression coefficient often shows the feature of the product category.

6. CONCLUSION AND FUTURE WORK

In this study, we aim to clarify the relationship between SNSs and purchasing on the EC site. Especially, in this paper we indicate the results of predictive models of purchasing behavior that was created using the information on Twitter. We focused on Twitter that is one of the most popular SNS. Moreover, we focused on the golf portal site that is the subject of this research. Firstly, we performed correlation analysis using the number of purchasing of product categories and the number of tweets of golf accounts. As the results, it became clear that "There is a weak positive correlation between the number of purchase and the number of tweets in 1 year" and "Tweets have strong positive correlation with a specific month, products and tweets."

Secondly, through the results of correlation analysis, we create 768 predictive models of purchasing behavior by logistic regression. We set the presence of customer's purchase data as objective variable for each visit to site. Additionally, we set marketing data and social data as explanatory variables.

Finally, we conducted hierarchical cluster analysis using parameter values of each models. As the result, it became clear that partial regression coefficient often shows

the feature of the product category.

In our future works, we will analyze the differences of these clusters. Moreover, we will analyze about the contents of tweets.

REFERENCES

Gupta, S. (1988) Impact of sales promotion on when, what and how much to buy, *Journal of Marketing Research*, **25**, 342-355.

Shaw, M.J., Subramaniam, C. Tan, G.W. and Welge, M.E. (2001) Knowledge management and data mining for marketing, *Decision Support Systems*, **31(1)**, 127-137.

Moe, W.W. and Fader, P.S. (2004) Dynamic conversion behavior at e-commerce sites, *Management Science*, **50(3)**, 326-335.

Koike, Y., Sugaya, K., Sumita, U. Takahashi, K., Hirano, T. and Yamamoto, K. (2007) "Real time distinction" of the important customers based on access log data, *Department of Social Systems and Management Discussion Paper Series*, 1177, 1-7. (In Japanese)

Hisamatsu, T., Togawa, T., Asahi, Y. and Namatame, T. (2013) A proposal of the purchase indication discovery model in EC site, *Communications of the Operations Research Society of Japan*, **58(2)**, 93-100. (In Japanese)

Liu, Y. (2006) World of mouth for movies: its dynamics and impact on box office revenue, *Journal of Marketing*, **70**, 74-89.

Mishne, G. and Glance, N. (2006) Predicting movie sales from blogger sentiment, in *AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 155-158.

Yoshida, N. Ishii, A. and Arakaki, H. (2007) *Equation of the smashing success: mathematize the personal influence effect of social media*, Discover 21 Inc. (in Japanese)

Tsurumi, H., Masuda, J. and Nakayama, A. (2013) Relevance analysis of the communication on twitter about goods and sales performance *Communications of the Operations Research Society of Japan*, **58(8)**, 436-441.

Tsurumi, H., Masuda, J. and Nakayama, A. (2015) Possibility and limitation of the text data on SNS utilization in marketing, *Marketing Journal*, **35(2)**, 38-54. (In Japanese)