Exploring customer needs of Huis Ten Bosch per customer attribute: Market segmentation and targeting by using blog text mining and conjoint analysis

Junichi Kato  
Tsukuba International University  
Tsuchiura, Ibaraki Prefecture, Japan  
kato0119@tius.ac.jp

Tetsuro Hyodo  
Tokyo University of Marine Science and Technology  
Tokyo, Japan  
hyodo@kaiyodai.ac.jp

Mamoru Imanishi  
Fukuoka University  
Fukuoka, Japan  
imanishi@econ.fukuoka-u.ac.jp

Saburo Saito  
Fukuoka University  
Fukuoka, Japan  
saito@econ.fukuoka-u.ac.jp

ABSTRACT

This study aims to determine customer needs per customer attribute through market segmentation and targeting by taking Huis Ten Bosch as one example. Conjoint analysis and RFM analysis show us that HTB customers of different customer attribute attach importance to different needs. We interpret that these needs can be satisfied by providing matching products and services. These customers who are high level or low level of R, F, M, or all RFM have different willingness to pay for each product attribute. By clarifying these different needs per customer attribute, managers can decide not to provide similar services for all customers and managers can decide which products or services are better suited for particular customers. Managers can focus only on important customers and attract them.

JEL Classification: L1, L71

Keywords: marketing, text mining, conjoint analysis, KIP, Huis Ten Bosch
INTRODUCTION

It is one of the most important management issues for an organization, which faces severe business environment to grasp the important customer needs by using targeting and attract customers who share similar needs by using market segmentation.

The aim of this paper is to clarify the customer needs per customer attribute through executing market segmentation and targeting by taking Huis Ten Bosch (HTB) as one example. This research has two objectives. The first objective is to clarify customer needs by using blog text mining. Blogs are the media which give us raw customers' voices. Blogs give us more massive natural language data than questionnaires. In this research, we disclose customer needs by using blog text mining.

The second object is to detect the more important needs per customer attribute. Second object includes two analyses. The first is Recency, Frequency, and Monetary (RFM) analysis. By using RFM analysis, customers are ranked from the viewpoint of Recency (customers recently visit HTB), Frequency (customers frequently visit HTB), and Monetary (customers spend more money for the trip to HTB or in HTB). These scores are customer attributes. The second is to grasp the important customer needs per customer attribute and price customers would like to pay to satisfy their needs by using conjoint analysis. These analyses enable us not only to execute the market segmentation but also to do targeting.

More detail outline is as follows. First, we use blog texts as data and execute the market segmentation by following Kato and Ishikawa Procedure (KIP). The results of this analysis were earlier presented in a non-refereed Japanese national conference. KIP is the procedure which shows us the customer needs from blog texts data by using text mining. It was difficult to interpret customer needs from results of analysis. Accordingly, we use the principal component analysis for this research. The results are shown in the section of results of blog text mining.

The advantages to use blog texts data for marketing research were already pointed out by existing marketing studies. Volo (2010) points out the following specifications of blog text data. We can use data which blog authors wrote spontaneously. Blog texts as data have less possibility that force test subjects to create data. Mitamura et al. (2008) use computer programs for analysis and process the big data we cannot analyze by hands. This means we can use massive natural language data.

Pan, MacLaurin, and Crotes (2007), Volo (2010), and Mitamura et al. (2008) employ a web search engine. These previous researches could collect various opinions as data from a wide range of subjects. The distinction of KIP is not from the viewpoint of an engineer but from the viewpoint of the previous marketing studies. KIP follows the definition of market segmentation that Kotler and Keller (2006) use.

Next, after getting the results of principal components analysis, we execute both RFM analysis and conjoint analysis. We can grasp important customer needs for managers by using both these methods. RFM analysis enables us to isolate customers who visit HTB recently, visit HTB frequently, or spend more money
when visiting HTB and in HTB. We clarify the needs which the customers value by using conjoint analysis. The results are shown in the section of results of RFM and conjoint analysis.

From the above results of blog text mining and results of RFM and conjoint analysis, needs about HTB which customers value are clarified per customer attribute. This procedure including KIP, RFM analysis, and conjoint analysis enables us to do targeting which could not be executed based on the results of past blog text mining alone. At this point, this research contributes to both academic research and business practices.

**METHODS**

**Blog text mining by using KIP**

The procedure of market segmentation from blog texts is shown in Figure 1. This is the procedure in Kato and Ishikawa (2011a). In this research, this procedure is called KIP. However, this research uses step 5 (Author Analysis) for step 4 or 5 in the procedure for author clustering. Additionally, this research executes principal component analysis in step 6 for easy labeling and easy interpreting of the results. Kato (2013) explains the procedure which used the principal component analysis in KIP in detail.

![Figure 1. Procedure of market segmentation](Image)

**Data for KIP is prepared as follows. First, researchers decide one word (Target Keyword) which expresses the market they would like to analyze and they...**
collect all goo blog authors who used the word (Target Keyword) even if for one time only. Next, researchers gather all blog texts of all blog authors who used the Target Keyword. Finally, researchers extract words from all blog texts and prepare words for analysis. This is row data for KIP. This step is data collection.

The next step involves selection of words which researchers use for clustering of blog authors. These words are used as criteria for clustering. Blog authors with the same frequency of use of selected words are clustered in the same group.

KIP uses two types of words as criteria for grouping of blog authors. One of these is words which pertain to the product. These are product keywords. The other type of words is personal keywords which denote the customers. Researchers choose these two types of words from row data and use these as criteria.

Blog authors who have the same frequency of use of these product keywords and personal keywords are gathered, with these keywords as criteria. These groups are the results of market segmentation.

Lastly, principal axes which comprehend each group are extracted and are put on labels. Because researchers cannot understand and interpret the needs which these groups share, principal axes which grasp each group of blog authors are extracted by using principal component analysis. And then, the labels are put on the basis of the eigenvector of each word for easy interpretation.

**Experimental Design by Using AlgDesign and Making of Web Questionnaire**

Questionnaires for conjoint analysis are made from principal axes which are extracted by using blog text mining. First, the experiment is designed. This research uses the AlgDesign package of the statistical environment R for the experimental design. This procedure is introduced in Aizaki and Nishimura (2007). According to Aizaki and Nishimura (2007), questionnaires are made by using six steps described in Figure 2.

Questionnaires are created by following this procedure. In this process, M in step 3 and 4 is 2, and P in step 5 is 8. These questionnaires are made by using two axes that are extracted from loyal authors and two axes that are extracted from longtail authors. The six steps are executed for loyal and longtail authors separately. As a result, 16 question items are made.

Lastly, the php program is made for investigating these items by web questionnaires. The web questionnaires are shown on a Web server. The php program reads the file of the items which is prepared, and generates the html file dynamically. When a subject replies to items, the answers are saved as raw data in a Comma Separated Value form (CSV) after having checked required items.

As a consequence of using undergraduate students as subjects, most of them usually do not have device (e.g., smart phone) to access the Internet and to see web questionnaires. Hence, the paper-based questionnaires which are completely the same as the web questionnaires are prepared, and the answers to the paper-based questionnaires are inputted by using web questionnaires. In other words, web questionnaires are used as the data input form. The web questionnaires are shown publicly.
RFM Analysis

Data gathered are analyzed by using RFM framework. The first step is data cleaning. Data are checked from the viewpoints of the same person’s repetitive answers, inappropriate answers, or theoretically impossible answers, and so on. The inappropriate answers are deleted from data.

The second step is allocation of RFM scores. The question items in this research are on the basis of the framework of Recency Frequency and Monetary (RFM). The data which are only in conjunction with these RFM are analyzed here. Researches on RFM framework are reviewed by Wei, Lin, and Wu (2010) in detail. The analysis of this paper is executed while referring to Wei, Lin, and Wu (2010).

First, data which are related to RFM are extracted from all data. Those data are Questions No. 5 to 11. Question No. 8 is about whether the subject has the experience of visiting HTB.

Second, these extracted data are allocated RFM scores. Recency is related to two types of questions. One is how many months have passed from the last trip, while the other is how many months have passed from the last visit to HTB. The smaller the number of months is, the higher scores of Recency are allocated. The 4 is allocated as score to subjects of the highest 25 percent of number of months. Then the 3, 2, or 1 are allocated per each 25 percent. This 25 percent denotes a percentile. Each customer is allocated one of scores from 4 to 1 as Recency score. Therefore, the higher the score is, the higher the loyalty is.

Frequency is related to two types of questions. One is how many times the subject has traveled in the recent five years. It includes similar places to which the
subject visits. The other is how many times the subject has visited HTB in the last five years. The 4 is allocated as score to subjects of the highest 25 percent. Then the 3, 2, or 1 are allocated for each 25 percent. This 25 percent refers to a percentile. Each customer is allocated one of scores from 4 to 1 as Frequency score. The higher the score is, the higher the loyalty is.

Monetary is related to two types of questions. One is how much subject spends for one travel. The traveling includes international and national trips. The cost for traveling includes transport and lodging expenses. The other is how much subject spends for one travel to HTB. The 4 is allocated as score to subjects of the highest 25 percent. Then the 3, 2, or 1 are allocated for each 25 percent. This 25 percent pertains to a percentile. Each customer is allocated one of scores from 4 to 1 as Monetary score. The higher the score is, the higher the loyalty is.

The third step is the narrowing of data. In the previous procedure, RFM scores and answers for questions from No. 5 to 11 are prepared. Data are more refined from these data. Only subjects who are allocated 4 or 1 as score of R, F, and M are selected. According to this third step, the following subjects’ sets are prepared. The subjects’ set as “q5-R1” includes the subjects who have 1 as score of R for question No.5. The “q6-F1” means the set of subjects who have 1 as score of F for question No.6. The “q7-M1” is the set of subjects who have 1 as score of M for question No.7. Similarly, the “q5-R4” is set of subjects who have 4 as the score of R for question No.5. The “q6-F4” is the set of subjects who have 4 as the score of F for question No.6. The “q7-M4” is the set of subjects who have 4 as score of M for question No.7. By using these data, the next RFM analysis can be executed.

The fourth step is an implementation of RFM analysis by using the narrowed-down data. In RFM analysis, the mean values are calculated by using each data of q5-R1, q6-F1, q7-M1, q5-R4, q6-F4, and q7-M4. For instance, in the set of subjects (q5-R1), the average of data of question No.5 is calculated. These mean values are the result of RFM analysis.

**Conjoint Analysis**

Conjoint analysis consists of three steps. Step 1 involves the preparation of data sets from questions No.12 to No.19. First of step1, subjects for conjoint analysis are only persons who have previously visited HTB. Therefore, data sets, answers of which are from questions No. 12 to No.19 are selected by the condition that subjects choose 1 in question No.8.

Next of step 1, items which are from questions No.12 to No.19 are principal axes of loyal authors. So data sets are chosen by the criteria that the RFM score is either 3 or 4 in the RFM analysis. For example, when q9-R3 or 4 means that R score in question No.9 is 3 or 4, sets of subjects, whose RFM scores are q9-R3 or 4, q10-F3or4, q11-M3 or 4, and q9 to 11-RFM3 or 4, are prepared.

Final of step 1, answers to questions from No.12 to No.19 are shaped into data sets, which can be analyzed by statistical environment R for conjoint analysis. The value (1 or 0), which corresponds to whether subject has each attribute or not, is inputted under the column headers. The column headers are price.planA, price.planB, price.non, medinfo.planA, medinfo.planB, medinfo.non, resort.planA, resort.planB, and resort.non. The data sets are prepared by using eight rows for
Each subject. The number of rows (eight rows) is the same as the number of questions (questions No.12 to No.19). These data sets are prepared for the number of subjects.

Table 1. Data set for conjoint analysis

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q12</td>
<td>A</td>
<td>6000</td>
<td>6500</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q13</td>
<td>B</td>
<td>5500</td>
<td>5000</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q14</td>
<td>B</td>
<td>5500</td>
<td>5500</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q15</td>
<td>N</td>
<td>5000</td>
<td>6500</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q16</td>
<td>N</td>
<td>6000</td>
<td>6000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q17</td>
<td>N</td>
<td>5500</td>
<td>5000</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Q18</td>
<td>N</td>
<td>5000</td>
<td>5500</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q19</td>
<td>N</td>
<td>6000</td>
<td>6500</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Q20</td>
<td>N</td>
<td>6000</td>
<td>6500</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Q21</td>
<td>N</td>
<td>6000</td>
<td>6500</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Q22</td>
<td>N</td>
<td>6000</td>
<td>6500</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Source: Authors

The word “mode” is added to the first part of column headers of the data sets which are prepared for the number of subjects. This column header means the plan which a subject chooses in questions from No.12 to No.19 (plan A, plan B, or neither). The words “sex”, “R”, “F”, and “M” are added to the last part of column headers. Then, the values which correspond to each column header are inputted. These values are repeated by the number of questions. If the number of questions is eight, the values are repeated eight times by using eight rows. Data sets are then prepared for conjoint analysis. A part of data sets is shown in the Table 1 as one example. The values in Table 1 are not real data. Additionally, expressions of column headers are simplified in this example. The “A” means “plan A”; “B”, “plan B”; “N”, “none”; “mi”, “medinfo”; and “re”, “resort”.

Step 2 requires the preparation of data sets from question No.20 to No.27. First of step 2, subjects of data sets are only persons who have previously visited HTB. Therefore, only subjects who choose 1 for question No.8 are selected. Their answers from question No.20 to No.27 are data sets.

Next of step 2, items which are questions No. 20 to No. 27 are principal axes of longtail authors. Data sets are extracted by the criteria that RFM scores are only 1 or 2. When q9-R1or2 means set of subjects who choose 1 or 2 in question No. 9 as R score, the q9-R1or2, q10-F1or2, q11-M1or2, and q9to11-RFM1or2 are prepared. They are a set of subjects whose scores are low for conjoint analysis.

Final of step 2, answers to questions No. 20 to No. 27 are made into data sets which can be used for conjoint analysis in statistical environment R. The answers which correspond to price.planA, price.planB, price.non, eco.planA, eco.planB, eco.non, blog.planA, blog.planB, and blog.non are inputted as the value 1 or 0 in each column. These answers (1 or 0) mean whether subject agrees with each attribute or not. The data sets are created by using eight rows which are the same as the number of questions (No. 20 to No. 27). These data sets are prepared for the number of subjects.

The word “mode” is added to the first part of column headers of the data sets. This column header means a plan which a subject chooses from questions.
No. 20 to No. 27 (plan A, plan B, or neither). The words “sex”, “R”, “F”, and “M” are added to the last part of column headers. Then, the values which correspond to each column header are inputted. These values are repeated by the number of questions. When the number of question is eight because the questions are from No. 20 to No. 27, the values are repeated eight times by using eight rows. Data sets are then prepared for conjoint analysis. The outline is the same as Table 1.

Step 3 is to execute conjoint analysis for each RFM score. Sets of subjects are extracted in step 1 and 2. These data sets are analyzed by using mlogit library of statistical environment R. The mlogit library enables users to build various types of the discrete choice models. These models include MNL (Multinomial Logit Model), NL (Nested Logit Model), MNP (Multinomial Probit Model), and MXL (Mixed Logit Model).

Data sets which are used in this research do not need to pay attention to the error structures. MNL of mlogit library is employed as a method of conjoint analysis. As is evident from Table 1, the number of choices is three in this model. For example, the probability of choices is expressed as the equation (1). The $\theta_k$ means parameters in the equation (2). $X_{ik}$ is kth explanatory variable (in this research, it is price, medinfo, or resort etc.) of alternatives i (in this research, it is plan A, plan B, or non).

\[
P_{planA} = \frac{e^{x_k V_{planA}}}{e^{x_k V_{planA}} + e^{x_k V_{planB}} + e^{x_k V_{non}}} \tag{1}
\]

\[
V_i = \sum_{k} \theta_k X_{ik} \tag{2}
\]

**RESULTS**

**Blog Text Mining**

In the first step, data are gathered by using target keyword “Huis Ten Bosch”. Blog sites are searched by using goo blog search engine. Nine hundred seven blog entries are retrieved. Authors of all these blog entries are gathered, total of which is 522. The time period of the collected data is from March, 9, 2004 to September, 18, 2011. Some dates of gathered blogs are unrealistic. For example, these years are 0000 or 1968. In this research, time periods is decided as follows.

The starting day of a goo blog site is March, 9th, 2004 (I checked http://pr.goo.ne.jp/detail/290/ on October 3, 2011). After checking dates of the retrieved data, the period of blogs are decided between that starting day of a blog site and the day at which data are gathered. Needless to say, the unrealistic dates of the gathered blogs do not mean that data of blogs are nonsense. So these blogs are still used as data in this research.

Blog texts are extracted from the blog entries. These blog texts are analyzed by morphological analysis. Words are drawn from these blog texts. If the retrieval rate of 90 percent cannot be achieved for a particular blog, the author of this blog is omitted from the list of 522 authors. There is not a rigorous criterion of acceptance of blogs as data. Over 90 percent is the figure used as a criterion in
this research, but over 99 percent or 100 percent is used as a criterion in the previous research. A rigorous criterion cannot be determined because the number of blogs which can be gotten depends on the condition of networks. Finally, 487 authors, 517,126 blog entries, and 914,109 words are shortlisted.

In the second step, product keywords are chosen. Product keywords are defined as similar words with target keyword “Huis Ten Bosch”. Words which are more similar with Huis Ten Bosch are employed as product keywords. The 10,230 words are selected as product keywords. In the third step, personal keywords are chosen. First, blog entries are gathered for each author. The words that one author uses but other authors do not use very much are defined as personal keywords. The 10,230 words are selected as personal keywords. The same number of words is used as product keywords because there is no criterion.

The fourth step is author analysis. An author analysis means categorization of authors per keyword. That categorization includes loyalty to a product. First, blog entries are clustered for each author. Then, these authors are categorized by product keywords that are selected in step 2. Each categorization forms groups of authors who used similar words as product keywords.

A relative ratio of significant scores of product keywords is calculated for each author of each cluster in these four clusters, and then mean scores of the 25 percent of upper side are calculated for each cluster. The cluster which has a maximum value of mean scores is defined as loyal authors. They have a high level loyalty to the product. The number of loyal authors is 104. In contrast, authors who have the minimum value of loyalty score are defined as longtail authors. The number of longtail authors is 116. Authors categorized by product keywords are grouped by personal keywords that were selected in step 3. This categorization creates four groups from 104 loyal authors and 116 longtail authors.

The fifth step is labeling. Four groups of loyal authors include 79,465 words (18 authors), 107,624 words (64 authors), 142,748 words (14 authors), and 32,168 words (eight authors). Four groups of longtail authors include 229,550 words (33 authors), 255,159 words (61 authors), 80,253 words (10 authors), and 343,445 words (12 authors). Four groups of loyal authors are unified to one group and four groups of longtail authors are unified to another. Loyal authors have 243,382 words and longtail authors have 672,155 words.

The values of significance of each word are calculated by using p-value of chi-squared. The smaller the p-value is, the smaller the probability of independence is. A small p-value of a word shows that the word characterizes the loyal authors or longtail authors. Meanwhile, chi-squared value of each word is calculated per each cluster for a later analysis. More important words that characterize authors than other words are chosen. Words with smaller p-value are selected as a criterion. The selected words are 0.1 percent of all words (calculating to one decimal place and rounding down if the number is less than 5). Results show that 243 words are selected from 243,382 words of loyal authors, and 672 words from 672,155 words of longtail authors.

A principal component analysis is executed for 243 words and 672 words by using values of chi-squared as data. Loyal and longtail authors consist of four clusters each. Four principal components that characterize these clusters are extracted by principal component analysis. From the result of principal component
analysis, the first and second principal components capture 85.9 percent of cumulative proportion in loyal authors and 78.8 percent of cumulative proportion in longtail authors. Almost 80 percent of loyal and longtail authors are explained by the first and second principal components. Therefore, the first and second principal components are employed for interpreting the needs of loyal and longtail authors.

These two components are interpreted as follows. The word with the maximum absolute value of eigenvector in the first and second component is found out. If the value of eigenvector is positive, words are sorted and 50 words are selected in descending order. On the other hand, if the value of eigenvector is negative, words are sorted and 50 words are selected in ascending order. These 50 words are used for interpreting customer needs.

Loyal authors have two axes. The first principal axis of loyal authors includes the following words: “www, http, html, jp, PDF, youtube, cigarette, research institute, new type, virus, information, specimen, influenza, blogspot, public welfare, US, Ministry of Labor, cause of disease”, and so on. These consist of information-related words and medical-related words. Hence, we put medical information as the label on the first principal axis of loyal authors. The second principal axis of loyal authors is characterized by the following words: “aviation, Tokyo, Chiba, Narita, Monaco, resort, Monte Carlo, hotel”, and so on. These are related to resorts. Therefore, we put foreign resorts as label on the second principal axis. Therefore, labels of loyal authors consist of medical information and foreign resorts.

On the other hand, longtail authors have two axes. The first principal axis of longtail authors includes: “diary, archives, seesaa, sale, SeesaaBlog, blog, umbrella, size, doll, html, livedoor, entry, FC, Blog”, and so on. Most of these words are related to blogs. Thus, we put blogs as label on the first principal axis. The second principal axis of longtail authors is characterized by the following words: “political party, last year, bank, finance, Diet, natural disaster, United States of America, the House of Representatives, March, committee, case, administration, prime minister, people, policy, budget”, and so on. These words are related to political and economic issues. Hence, we put political and economic issues as label on the second axis of longtail authors. Labels of longtail authors consist of blogs and political and economic issues.

Algorithm Design and Data Collection

Web questionnaire is created for clarifying the degree of the effect of customer choice probability on the basis of the principal axes. However, the paper-based questionnaire is used because most subjects cannot access the website. Web questionnaire is used for the data input format in this research.

Step 1, the result of question items which are based on the principal axes of loyal authors is shown. First of step 1, two axes are extracted as attributes of loyal authors from the results of principal component analysis. These two axes are medical information and accommodations like foreign resorts as shown in Table 2. These two axes have two levels each. Additionally, four levels of price are set by referring to the real prices of HTB.
Table 2. Factors and levels from loyal authors

<table>
<thead>
<tr>
<th>Loyal</th>
<th>Providing medical information from HTB</th>
<th>1.YES</th>
<th>2.NO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Having facilities like foreign resorts in HTB</td>
<td>1.YES</td>
<td>2.NO</td>
</tr>
<tr>
<td></td>
<td>Price (One day entrance fee including the fee of some institutions)</td>
<td>1.6500JPY</td>
<td>2.6000JPY</td>
</tr>
</tbody>
</table>

Source: Authors

Second of step 1, following six steps of Aizaki and Nishimura (2007) which is shown in method part of this paper, combinations of these attributes and levels are prepared by using AlgDesign library. Table 3 simply shows the results of attributes and levels which are created by using AlgDesign. The values in Table 3 are numbers which are allocated to each attribute of Table 2. Question items (questions No. 12 to No. 19) are prepared from axes of loyal authors on the basis of Table 3.

Table 3. Combinations of factors and levels by using AlgDesign from loyal authors

<table>
<thead>
<tr>
<th></th>
<th>Plan A</th>
<th></th>
<th>Plan B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 12</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Q 13</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Q 14</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Q 15</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Q 16</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Q 17</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Q 18</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Q 19</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Source: Authors

Table 4. Factors and levels from longtail authors

<table>
<thead>
<tr>
<th>Longtail</th>
<th>Providing information by using blogs from HTB</th>
<th>1.YES</th>
<th>2.NO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holding events to learn political and economic issues</td>
<td>1.YES</td>
<td>2.NO</td>
</tr>
<tr>
<td></td>
<td>Price (One day entrance fee including the fee of some institutions)</td>
<td>1.6500JPY</td>
<td>2.6000JPY</td>
</tr>
</tbody>
</table>

Source: Authors

Step 2, question items, which are based on longtail authors are shown. First of step 2, two axes are extracted as attributes of longtail authors from the results of principal component analysis. The two axes are the sending and receiving information by using blogs and the holding of educational events about political
and economic issues in Table 4. Each axis has two levels. Four levels prices are set by referring to the real prices of HTB.

Second of step 2, question items are prepared by combining attributes and levels in Table 4. The method to prepare these by using AlgDesign library is summarized as six steps of Aizaki and Nishimura (2007). Table 5 simply shows that the results of combination of attributes and levels are created by using AlgDesign. The values in Table 5 are numbers which are allocated to each attribute of Table 4. From Table 5, question items (questions No. 20 to No. 27) are created on the basis of axes which are extracted from longtail authors. Conjoint cards are created from the results of loyal and longtail authors.

Table 5. Combinations of factors and levels by using AlgDesign from longtail authors

<table>
<thead>
<tr>
<th></th>
<th>Plan A</th>
<th></th>
<th>Plan B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PRICE</td>
<td>BLOG</td>
<td>ECON</td>
<td>PRICE</td>
<td>BLOG</td>
</tr>
<tr>
<td>Q 20</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Q 21</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Q 22</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Q 23</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Q 24</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Q 25</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Q 26</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Q 27</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Source: Authors

Questionnaire is prepared by combining this conjoint card and question items of RFM. The affiliations of test subjects are Nakamura Gakuen University (Total 415 response includes the 288 valid response.), Tsukuba International University (Total 50 response includes the 45 valid response.), and Fukuoka University (Total 33 response includes the 28 valid response.). The study period is from July 9 to July 18, 2013.

RFM and Conjoint Analysis

RFM. Not only are subjects questioned about the experience of visiting HTB, but also about each RFM. The ratio of visiting HTB is 45.7 percent. The RFM analysis is in accordance with the procedure which is explained in the method section in this paper.

The subjects, whose Recency score is 4, let pass (1) 1.49 months on average from the day of last trip to the day of answering the questionnaire, and (2) 3.65 months on average from the day of last visit to HTB to the day of answering the questionnaire. On the other hand, the subjects, whose Recency score is 1, let pass (1) 20.06 months on average from the day of last trip to the day of answering the questionnaire, and (2) 104.83 months on average from the day of last visit to HTB to the day of answering the questionnaire.

The subjects, whose Frequency score is 4, travel to (1) places which include the same destinations 10.20 times during last five years, and (2) HTB 1.51 times during the last five years. On the other hand, the subjects, whose Frequency score
is 1, (1) travel to places which include the same destination 0.91 times during the last five years, and (2) have not traveled to HTB during the last five years. This means that there are no subjects who have smaller frequency than 25 percentile point.

The subjects, whose Monetary score is 4, spend (1) 70,994.87 yen on average per one time traveling which includes travel expenses and the cost of staying, and (2) 40234.38 yen on average per one trip to HTB which includes traveling expenses and the cost of staying. On the other hand, the subjects, whose Monetary score is 1, spend (1) 17031.75 yen on average per one trip which includes travel expenses and the cost of staying and (2) 4450 yen on average per one trip to HTB which includes travel expenses and the cost of staying.

**RFM and conjoint analysis.** Conjoint cards consist of (1) two principal factors with two levels each from loyal authors and four levels of prices, and (2) two principal factors with two levels each from longtail authors and four levels of prices. Answers of subjects with high-level score in each RFM to conjoint cards made from loyal authors’ axes are analyzed.

For the results, the subjects of 3 or 4 in Recency scores have the following values. The estimated coefficient value of price is -0.00092 and its probability of significance is 0.1 percent. The estimated coefficient value of foreign resort is 1.00224 and its probability of significance is 0.1 percent. The estimated coefficient value of medical information is 0.38682 and its probability of significance is 1 percent. The Willingness to Pay (WTP) of each factor is 1,089.39 yen for foreign resort and 420.46 yen for medical information.

The subjects of 3 or 4 in Frequency scores have the following values. The estimated coefficient value of price is -0.00086 and its probability of significance is 0.1 percent. The estimated coefficient value of foreign resort is 0.96832 and its probability of significance is 0.1 percent. The estimated coefficient value of medical information is 0.50451 and its probability of significance is 0.1 percent. WTP of each factor is 1,125.95 yen for foreign resort and 586.64 yen for medical information.

The subjects of 3 or 4 in Monetary scores have the following values. The estimated coefficient value of price is -0.00079 and its probability of significance is 0.1 percent. The estimated coefficient value of foreign resort is 1.02563 and its probability of significance is 0.1 percent. The estimated coefficient value of medical information is 0.58431 and its probability of significance is 0.1 percent. WTP of each factor is 1,298.27 yen for foreign resort and 739.63 yen for medical information.

The subjects of 3 or 4 in all Recency, Frequency and Monetary scores have the following values. The estimated coefficient value of price is -0.00098 and its probability of significance is 0.1 percent. The estimated coefficient value of foreign resort is 0.99357 and its probability of significance is 0.1 percent. The estimated coefficient value of medical information is 0.34878 and its probability of significance is 5 percent. WTP of each factor is 1,013.85 yen for foreign resort and 355.90 yen for medical information. The results are summarized as follows.

Lastly, answers of subjects of low score in each RFM to conjoint cards made from longtail authors’ axes are analyzed. For the results, the subjects, whose Recency score is 1 or 2, have the following values. The estimated coefficient value
of price is -0.00079 and its probability of significance is 0.1 percent. The estimated coefficient value of blogs is 0.61439 and its probability of significance is 0.1 percent. WTP of blogs is 777.71 yen.

Table 6. Conjoint analysis high level of RFM

|        | Estimate | Pr(>|t|)        | Log-Likelihood | McFadden R^2 | Likelihood ratio test  |
|--------|----------|----------------|----------------|--------------|------------------------|
| R 3 or 4 | PRICE   | -0.00091793 | 2.731e-14 ***  | -639.54      | 0.11101                |
|        | RESORT  | 1.00223851  | 3.819e-14 ***  |              |                        |
|        | MEDINFO | 0.38681586  | 0.001928 **    |              |                        |
| F 3 or 4 | PRICE  | -0.00085805 | < 2.2e-16 ***  | -875.83      | 0.1015                 |
|        | RESORT | 0.96831503  | < 2.2e-16 ***  |              |                        |
|        | MEDINFO| 0.50451261  | 2.549e-06 ***  |              |                        |
| M 3 or 4 | PRICE  | -0.00079324 | 3.553e-14 ***  | -819.57      | 0.1031                 |
|        | RESORT | 1.02562708  | < 2.2e-16 ***  |              |                        |
|        | MEDINFO| 0.58430942  | 5.451e-08 ***  |              |                        |
| RFM    | PRICE  | -0.00097724 | 3.645e-10 ***  | -368.12      | 0.11645                |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: Authors

The subjects, whose Frequency score is 1 or 2, have the following values. The estimated coefficient value of price is -0.00084 and its probability of significance is 0.1 percent. The estimated coefficient value of blogs is 0.45182 and its probability of significance is 0.1 percent. WTP of blogs is 537.88 yen.

Table 7. Conjoint analysis low level of RFM

|        | Estimate | Pr(>|t|)        | Log-Likelihood | McFadden R^2 | Likelihood ratio test  |
|--------|----------|----------------|----------------|--------------|------------------------|
| R 1 or 2 | PRICE   | -0.00079346 | 7.055e-11 ***  | -625.84      | 0.076494               |
|        | ECON    | 0.11594044  | 0.3809387     |              |                        |
|        | BLOG    | 0.61438882  | 2.965e-06 ***  |              |                        |
| F 1 or 2 | PRICE  | -0.00080841 | 1.973e-07 ***  | -392.56      | 0.089206               |
|        | ECON    | 0.18152163  | 0.2871047     |              |                        |
|        | BLOG    | 0.72952873  | 1.896e-05 ***  |              |                        |
| M 1 or 2 | PRICE  | -0.00084387 | 1.021e-08 ***  | -440.38      | 0.05992                |
|        | ECON    | -0.03079431 | 0.848495      |              |                        |
|        | BLOG    | 0.45181732  | 0.004951 **    |              |                        |
| RFM    | PRICE  | -0.00108307 | 8.918e-05 ***  | -120.4       | 0.15255                |
|        | ECON    | 0.17328520  | 0.562626      |              |                        |
|        | BLOG    | 0.69249649  | 0.019809 *     |              |                        |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: Authors

The subjects, whose all Recency, Frequency and Monetary scores are 1 or 2, have the following values. The estimated coefficient value of price is -0.00108 and its probability of significance is 0.1 percent. The estimated coefficient value of
blogs is 0.69250 and its probability of significance is 5 percent. WTP of blogs is 641.20 yen. The results are summarized as the tables.

**DISCUSSION AND CONCLUSION**

This research has two objectives. The first objective is to clarify customer needs by using blog text mining, while the second object is to detect the more important needs per customer attribute. Second object includes two analyses. The first is RFM analysis. By using RFM analysis, customers are ranked from the viewpoint of Recency, Frequency, and Monetary. These scores are customer attributes. The second is to grasp the important customer needs per customer attribute and price customers would like to pay to satisfy their needs by using conjoint analysis. These analyses enable us not only to execute the market segmentation but also to do targeting.

The results of extracting of customer needs from blog texts are already shown in the section of the results of blog text mining as interpretations of principal axes. The results of RFM and conjoint analysis are the heart of the discussion of this section. This discussion sheds light on important customer needs per customer attribute like RFM from the viewpoint of customers.

In this section, our discussion consists of two parts per customer attribute (RFM score). First, the customer needs which customers with the high RFM level value and price which customers would like to pay to satisfy their needs are analyzed. Second, the customer needs which customers with the low RFM level value and price which customers would like to pay to satisfy their needs are discussed. This procedure enables to execute targeting which the existing blog text mining has never been able to execute.

**Customers of high RFM level**

When customers have scores of 3 or 4 on Recency, Frequency, Monetary, or all RFM, the factors like foreign resorts and medical information significantly affect the probabilities of choices. Therefore, foreign resorts and medical information are important factors for them.

However, the degrees of importance are different on the basis of the values of coefficients. Next, the differences of importance are emphasized. First, the coefficients of foreign resorts for customers whose scores in Recency, Frequency, Monetary, or all RFM are 3 or 4 are compared. The order by which tourists regard the facilities like foreign resorts as the most important factor is as follows. The first is customers who spend more money for the trip to HTB, followed by customers who recently visited HTB, then customers who recently did and frequently visit it and spend more money, and customers who frequently visit it.

Second, the coefficients of medical information for customers whose scores in R, F, M, or all RFM are 3 or 4 are compared. The order by which subjects regard medical information as the most important factor is as follows. The first is subjects who spend more money for the trip to HTB, followed by subjects who recently did visit HTB, then subjects who recently did visit HTB, followed by subjects who recently did and frequently visit HTB and spend more money for the trip to
HTB. These orders are one criterion for selection of priority of product attributes when attributes of target customers are decided.

The selection of product attributes depends on the balance between attributes and price. Customers who regard price as the most important are subjects whose scores of RFM are 3 or 4. Therefore, when target is these subjects, managers need to pay attention to the price. The clue to make decision about price is WTP.

Third, WTPs of subjects whose scores in Recency, Frequency, Monetary, or all RFM are 3 or 4 are compared. Subjects who spend more money think of spending the most money for facilities like foreign resorts. The price is 1,298 yen. Similarly, subjects who spend more money also think of spending the most money for medical information. The price is 740 yen. Subjects who spend more money would like to maintain facilities like foreign resorts, even if they pay 1,298 yen. They would like to be provided medical information even though they pay 740 yen.

On the contrary, the customers who have the least amount of WTP to facilities like foreign resorts and to medical information are subjects who recently visited and frequently visit HTB and spend more money. The prices of their WTPs are 1,014 yen to foreign resorts and 356 yen to medical information. They recently visited and frequently visit HTB and spend more money. It is assumed that such customers are critical of price. When managers provide the target customers with the products and services, they can refer to this WTP and can set a realistic price.

From these discussions, customers who spend more money for the trip to HTB regard facilities like foreign resorts and medical information as more important than customers with other attributes. Therefore, the price they would like to pay is the highest of all compared to customers with other attributes. If they are the target customers, facilities like foreign resorts and medical information are effective even though within the limited range of WTP.

On the other hand, customers who recently did and frequently visit HTB and spend more money for the trip to HTB do not regard medical information as important and do not think of paying much money for medical information when their attributes are compared with other attributes, even if their attributes significantly affect the probabilities of choice. To provide medical information for customers who have the same attributes as theirs is not as effective compared to those of other attributes.

Customers of low RFM level

Customers whose scores in Recency, Frequency, Monetary, or all RFM scores are 1 or 2 are discussed. They do not significantly regard events for learning about political and economic issues as important but regard providing information by using blogs as important. Therefore, providing information by using blogs is important for attracting them.

However, the degrees of importance are different. First, coefficients of blogs of customers whose Recency, Frequency, Monetary, or all RFM scores are 1 or 2 are estimated and compared. The order by which customers regard blogs as important is as follows. The first is customers who do not frequently visit HTB very much, followed by customers who did not recently and do not frequently visit HTB
and who do not spend more money for the travel to HTB, then customers who did not recently visit HTB, and customers who do not spend more money for the trip to HTB. Therefore, it seems that knowing the present situations of HTB by blogs is useful for customers who do not visit HTB very much.

The important product attributes for each customer attribute are criteria to decide the priority for selection of target customers. This is especially true when managers would like to expand the market, and customers who have the low-level scores become the target. The important product attributes for each customer attribute are keys. It is clarified that customers who do not frequently visit HTB very much evaluate providing information by using blogs as the important compared to customers with other attributes.

The selection of the important product attribute depends on the balance between attributes and price. Customers who regard price as the most important thing are subjects who did not recently and frequently visit HTB and who do not spend more money for the trip to HTB. This means that these customers have low level scores in all RFM. This is the rational and appropriate result. Therefore, when managers target these customers, they need to pay attention to price setting. The clue for managers to decide the price setting is WTP.

Second, WTPs of customers whose scores in Recency, Frequency, Monetary, or all RFM are 1 or 2 are compared. Customers who do not frequently visit HTB think of spending the most money for providing information by blogs. The price is 901 yen. On the contrary, customers who do not spend money for the trip to HTB very much have the least WTP for providing information by blogs. The price is 538 yen.

Therefore, the amount of money which customers who do not spend money very much would like to pay is calculated. This amount is about half of the amount of money which customers who do not frequently visit HTB would like to pay. This difference of WTP is clue for managers to decide the realistic price setting.

Customers who do not frequently visit HTB regard providing information as more important thing and think of paying more money than customers with other attributes. Therefore, when they are the targets for expanding the market, providing information by blogs is the effective product attribute even though the range is limited within WTP.

On the contrary, customers who do not frequently visit HTB do not regard providing information by blogs as more important than customers with other attributes. Therefore, providing information by blogs is not effective for customers who do not visit HTB very much compared to customers with other attributes, even if it significantly affects the probabilities of choice.

For each group of customers of high and low RFM level, the important product attributes and WTP per customer attributes are examined. Deciding the important product attribute and appropriate price setting for each target customer helps to create the plan for attracting customers. This procedure enables managers to execute targeting per customer attributes, which the existing blog text mining has never been able to execute. This is the contribution of this research.
Finally, the remaining future research issues are summarized. First, the samples of this research questionnaire are undergraduate students. Therefore, it is hard to generalize from the results of this research. Generalization of this research needs a rigorous sampling. Second, the php program needs to be revised for creating web questionnaire form dynamically. Lastly, some items of questionnaire are not encoded well by after-coding. In future researches, the way to collect data needs revision. These are the limitation of this research.

REFERENCES


