Internet Customer Segmentation Using Web Log Data

Jae Jeung Rho, (E-mail: jjrho@icu.ac.kr), Information & Communications University, South Korea
Byeong-Joon Moon, Kyung Hee University, South Korea
Yoon-Jeong Kim, Venture People Corp., South Korea
Dong-Hoon Yang, Information & Communications University, South Korea

Abstract

The objective of this paper is to analyze web transaction log data that reveal customer behavior in the Internet channel, and to provide a useful online customer segmentation scheme. To achieve this, we analyze the relationship between the behavior of customers for online pet shops and revenue. We use the decision-tree method as a data-mining technique, and clustering analysis to segment customers. We perform the study in two stages. First, we investigate the web transaction data of both the member customers and nonmember customers of a Korean online pet shop. Second, we narrow down the study focus and analyze only the member customers' demographic data and their web transaction data. As a result, we obtain several meaningful segments based on customers' transaction behavior and demographic characteristics. We use web log data to analyze customer transaction behavior and log-in information to analyze customer demographic characteristics. We discuss some strategic implications, for online shopping mall marketing, suggested by the acquired market segments.

Introduction

There are many skeptics who say that the Internet is nothing more than a new distribution channel, involving no special changes in customer behavior. In some sense, this is true. However, it is obvious that online markets provide rich new sources of data, which enable us to examine behavioral phenomena. In the past, these were impossible to study using more traditional sources. Leeflang and Wittink (2000) give a traditional marketing model dealing with grocery products that involve Universal Product Code (UPC) scanner data. The dataset concerns the mature stage of the product life cycle, and the data source is very rich. This research reveals valuable information for traditional firms (Mahajan, 2000). Although Leeflang and Wittink (2000) offer a valuable critique of marketing modeling of traditional markets, it should be complemented by modeling for Internet-driven products and activities. In this respect, there is another approach using web transaction data: web usage mining. In other cases, web log data can be used to find different patterns in customers' behavior, which can be harnessed for e-commerce marketing.

Levin's (2001) study addresses how consumers' Internet usage patterns influence the number of visits to a site, by using online panel data. Englis and Solomon (2000) describe a web-based data collection technique, suitable for a broad variety of consumer research, using participants' rapid-response feedback. These are two examples of the many studies regarding online customer behavior that have been empirically performed using online user surveys. As approaches to web transaction data become refined, supplying more detail than before, this data could be a good source for online marketing modeling instead of survey data.

The objective of this paper is to show customer segmentation based on web log data, including customer demographic data on category killers combined with online commerce and online communities. Since customers are not homogeneous in their preferences, wants, and needs, the idea of this research is to partition the market into groups or segments of "like" people, with similar needs and characteristics, who are likely to exhibit similar

purchasing behavior (Levin, 2001). Segmenting customers by given factors will reveal important and meaningful subgroups as well.

In this research, we use log data for short time periods, and connect it to other customer profile data such as customers' sex, education, age, and billing data. Data cubing conducted by this preprocess is more informative than log and profile data themselves. Using filtered and combined data, we finally divide customers into several meaningful subgroups through clustering analysis. We collected the web log data of an online retail shop that sells pet supplies in Korea and preprocessed and examined the customer segmentation. Like other retail shops, this online retailer uses a membership system for obtaining customer information. However, it also sells its products to nonmembers, who do not offer customer information and cannot take part in the online community offered in this shop. Accordingly, we examine two parts of the data set; the first dataset is full log access data ordered by accessing time, the other is members' log access data with customers' demographic data.

This paper also reveals a positive relationship between click stream data (e.g., visit frequency, duration per visit, clicks per visit) and purchasing propensity. Previous studies suggest that customers who shop frequently may be more likely to make a purchase on any given shopping occasion. However, we find that online retail shops with communities could be different in this respect, and we can show more detailed customer segmentation. We found that the full dataset is divided into four groups, and that the member dataset is divided into six groups with different characteristics. In marketing perspectives, this analysis can be informed by purpose of visit, ratio of each criterion, main criterion of classification, and its specification.

Conceptual Background

To understand the "real value" of customers, firms should model the marketing strategy using recency (time since the last purchase), purchase size, transaction frequency, and margin of products purchased. It is possible that the Internet provides a unique opportunity to observe how frequently visitors and customers come to a site, how long they remain, how they interact with content and tools, and how business is transacted.

In this respect, Hoffman and Novak (1999) prove that the "interactive metrics" of duration time and browsing depth, proposed to measure marketing effectiveness, are highly correlated with a compelling online experience. Furthermore, it is more important to understand the metrics of visitor retention in relation to the economic robustness of the site than simply to measure the number of visits.

To obtain information on web customer transaction data, many computer scientists have tried to find out how to measure customer behavior. Rosenstein (2000) details the possibilities and pitfalls of using web server logs to understand customer behavior on a web site. He considers the information recorded by the server and what legitimate inferences can be made from the data.

In another sense, marketing scientists try to model online marketing through empirical analysis based on web-based user surveys or offline surveys. Emmanouilides and Hammond (2000) explore the factors that predict Internet usage patterns through the use of consumer panel data.

Through the online user's telephone survey, Donthu (1999) provides insights on the Internet shopper. According to these results, Internet shoppers exhibit the following traits.

- Internet shoppers are older and have higher incomes than non-Internet shoppers.
- Internet shoppers are more convenience seeking, innovative, impulsive, and variety seeking, and less risk averse, than non-Internet shoppers.
- Internet shoppers are less brand and price conscious.
- Internet shoppers have a more positive attitude toward advertising and direct marketing than non-Internet shoppers.

However, in understanding customer behavior on the web, it is best segmented through web transaction data. A McKinsey marketing report shows a customer segmentation model developed by an online retailer, presenting the following table.

Table 1 shows the radically different behaviors and values of the segments, enabling the retailer to develop meaningful marketing action plans to improve profitability. This segmentation pertains to retail shops, which can be categorized into several groups according to scale of sales, selling product, and contents served.

Forrester Research (2000) further classifies retailers according to selling products, dividing them into three groups, as shown in Table 2.

The dataset used in this research is from online pet supplier that is a typical affinity-oriented retailer; in other words, a category killer. Another example of this type in Table 2, dELiAs.com, deals in casual apparel and related accessories for women between the ages of

Table 1 Understanding Key Customer Segments to Determine Potential Actions

	Avg. sale per item	Avg. item per order	Number of visits	Avg. visit duration
Loyalists	$60	1.7	12	15
Big Ticket Repeaters	$110	1.3	8	13
Explorers	$19	2.0	5	21
Targeted Item Seekers	$110	1.0	3	18

Table 2 Three Types of Retailer

	Type of Online Retailer		
	Broadline-based	**Activity-focused**	**Affinity-oriented**
Brand Authority	One-stop shopping	Available for all goods on specific theme	Available for goods on personal trend
Contents	Shallow	Deep	Deep
Target Customer	Mass marketing	One or more	One
Main Characteristics	Rating by customers Cross-sale-oriented	Special goods Counseling service	Mass-customized goods Membership service
Requirement for Investment	Very High	Average	Low
Category	Various	Various on certain category	Various on certain category
Example Site	Amazon.com Wal-Mart.com	Garden.com Cooking.com	SeniorNet.com dELiAs.com

10 and 24, and has recorded a 224 million dollar market capability, compared to Amazon.com's 5,656 million dollars as of July 5, 2002 (finance.yahoo.com). In general, a category killer is a small-sized firm and its requirements for investment are low.

The small-sized online retailers in Korea show clearer differentiations in business scale. According to the "Statistics on e-commerce in Korea 2002," the ratio of the total number of broadline e-tailers to category killers is 15:85, in contrast to the ratio of their revenues, 74:26. In this respect, category killer shops, combining commerce with community, need to use the synergy effect from both functions. Because they face excessive competition caused by low investment requirement, category killers may well make marketing efforts using such efficient means as online community services.

As for the relationship between virtual communities and their profitability, there are not many empirical researches. However, Kozinets (1999) defined virtual communities of consumption as a specific subgroup of virtual communities that explicitly center on consumption-related interests. Therefore, we regard the Korean pet shop site as a category killer in the realm of virtual communities, and we segment its customers within a marketing perspective.

Web usage mining

Web usage mining refers to the process of extracting marketing intelligence from a vast amount of web transaction data. The conglomerate of quasi-standardized log file formats requires a holistic approach to collecting, preprocessing, and consolidating available web site information in order to provide flexible, materialized views for explorative operations such as online analytical processing and data mining (Büchner, 1999). Modeling the user's behavior when navigating a web site is very relevant in e-commerce applications (Cooley *et al.*, 1999). User modeling first involves automatic segmentation of users displaying similar behavior, and automatic classification of users by means of identified user segments. The ultimate process of user modeling matches interactive web pages continuously adapted to user behavior.

Web log file

Web browsers and servers communicate using the stateless hypertext transfer protocol (HTTP). The header of an HTTP request message contains the attribute value pairs that a web server can record in its log file. Therefore, the web log file contains fields that describe each request that a browser makes from a server. By combining what can be inferred from this data, coupled with our understanding of how this information was derived, we can accurately analyze customer activity (Rosenstein, 2000).

Figure 1: An Example of Server Log Fields

Originating IP: 198.81.129.99
Timestamp: [26/Jul/1999:10:26:56 –0400]
HTTP Command & Protocol Version: "GET /ido/omages/id.gif HTTP/1.0"
Status Code: 200
Bytes Transferred: 660
Browser: "Mozilla/4.51 [en](WinNT; U)"
Referring URL: "http://www.company.com/"

There are two kinds of log file format: CLF (Common Log Format) and ECLF (Extended Common Log Format). According to a server manager's option, different contents can be accumulated. Figure 1 shows an example of server log fields.

Preprocessing

The first step in Web usage mining is preprocessing, which includes the domain-dependent tasks of data cleaning, user identification, session identification, and path completion. Data cleaning is the task of removing useless log entries. User identification is the process of matching page references with those who have records of profile. Session identification takes all of the page references for a given user in a log and breaks them up into user sessions (Cooley *et al.*, 1999). A session is the time duration of a visit, from the first access time to the last access time. In this stage, the site manager can acquire general usage statistics, such as number of 'hits', page most accessed per significant time period, and average time spent per day.

Web log data hypercube

In order to prepare a repository for further analytical activities, the next three steps are as follows (Büchner, 1999).

1. A web log data hypercube H represents n-dimensional information space $H = [D_1, D_2, D_3, D_4, .. D_n]$.
2. A dimension D represents m attributes $D = [a_1, a_2, a_3, a_4, .. a_m]$.
3. Total number of the cell is calculated as $|C| = \Pi|H_i|$, where $|H_i|$ is the cardinality of H (number of attributes).

Schematic design

For further analysis activities, a schema based on the relational calculus is modeled to represent relations between data. Each dimension is represented by a fact table that connects elements in a data model and summarization information (Büchner, 1999). An e-commerce fact table contains several key fields (i.e., CustomerKey, ProductKey, DataKey, SessionKey), including some statistical summarization information (e.g., Quantity, TotalPrice, ClickThroughRate), as shown Figure 2.

Customer segmentation

The basic concept of the segmentation is that markets are different from each other and are heterogeneous in terms of market need. The segmentation starts with the idea that a large group is composed of many subgroups. One of the primary purposes of segmentation is to identify subgroups that may be more profitably targeted than the mass market, according to customer lifetime value (CLV), from the point of view of a company. By tracing past customers' transaction data or behavior data, researchers can predict a probabilistic CLV of future customer transactions. In this respect, if the customer group were divided into smaller groups, and each subgroup had its own characteristics, a marketer could predict a CLV more accurately. Furthermore, Internet based commerce allows a greater customization, focusing on narrower targets with less cost (Weinstein, 2000). Empirical methods have been used for customer segmentation to determine which segment of customers a company should focus on attracting or retaining from the mixture of the profitable and the nonprofitable. Using past customer purchasing behavior, researchers can infer or calculate probabilities of purchase in the near future. Weinstein (2000) suggests two basic approaches for segmenting a user base as follows:

- *A priori*. This approach uses a hypothesis about the marketplace to divide it into segments. It breaks up markets along demographic, geographic, or behavioral lines.
- *Post hoc.* This approach allows the data to drive the discovery of consumer characteristics, predictive of responses to the product or marketing mix. Once key predictors have been identified, respondents are assigned to groups based on a set of rules. The process typically involves statistical methods.

E-metrics for segmentation

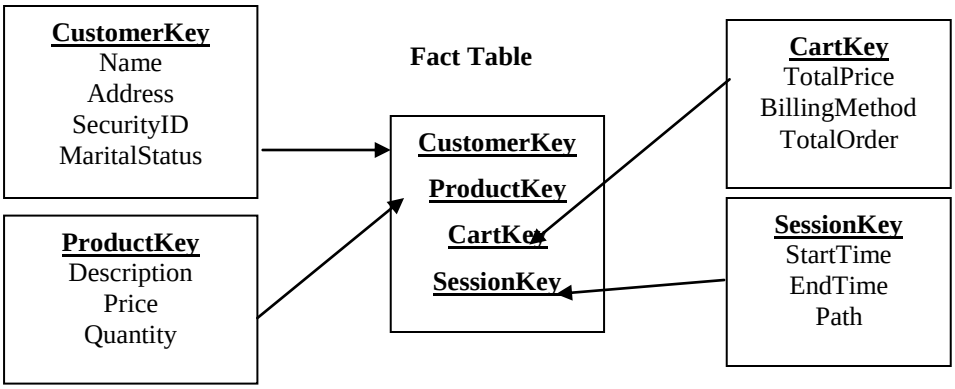
Even though the actual psychological processes that people experience are different, their behavioral patterns are strikingly similar (Fader, 2001). Online customers' behavioral patterns could be found by tracing online usage. One Internet consulting company, NetGenesis (www.netgen.com), proposes "e-metrics" that express meaningful ratios in various aspects. Ratios such as duration per session and number of pages viewed per session are said to be alternatives for investigating customers' behavioral patterns. According to NetGenesis, the meaningful e-metrics for inferring web users' activities, and terms on log analysis, are as shown in Table 3.

Problems

As discussed in the previous section, segmentation is performed for the purpose of identifying subgroups that may be profitably be targeted from the point of view of a company. We can consider the customer segmentation of general Korean online shopping malls from previous research. According to Song *et al.* (2001), online shopping mall users are divided into three subgroups with the characteristics shown in Table 4.

This study shows average purchase amount, visit frequency, and users' characteristics, for a general online market. However, this survey does not offer enough of an idea of what e-customers are, because the online shopping mall is divided into several subgroups with many differences such as sales scale, main sales point, main customers, and management style. For example, it is clear that an online shop selling computer hardware differs from a shop selling flowers or books in terms of price per head or distribution by sex, as well as in terms of distribution of customer transactions.

Figure 2 Web Log Schema



From a computer science perspective, there are trials for finding e-customer characteristics, using large volumes of sales histories, web transactions, and information data, to identify customer behavior. Hao (2001) has studied e-customer behavior using web log data, and has found that its commercial characteristics are not so different from those of the traditional market. He finds that the customers with the most sales usually visit more often and purchase more products, and that most customers are one-time visitors. We here perform our analysis within a marketing perspective, using data collected from Korean online retailers. Our research with online retailers that offer online communities is a first contribution in elaborating this perspective.

There have been many questions about how online communities affect the revenue of Internet firms. A potentially powerful way of organizing in the new Internet world is through the medium of the virtual Internet community. Rothaermel and Sugiyama (2001) stress that an effectively managed virtual community provides economic gains to the community organizer and to its members. This study theorizes on individual-level factors influencing an individual's economic transactions within a virtual Internet community, and then shifts the level of conceptualization to the community level and presents propositions with respect to the commercial success of virtual Internet communities.

Kozinets (1999) finds that on bulletin boards and in chat rooms, consumers form e-communities using networked computer technology to improve their knowledge, to socialize, to organize, and to play. There are multiple opportunities for marketers to insert, defend, alter, and reinforce brand meanings in all of these environments. However, virtual communities are difficult in some ways because they demand that marketers commit to the satisfaction and support of the community as well as the individual. Companies that do not do this may find that consumers with a strong need for community have migrated to a competitor that can offer access to and positive relations with an alternative or more desirable community. Yet, by following a membership or subscription strategy, insiders' knowledge and connections, and consequently elevated status in a meaningful and satisfying virtual community of consumption, can be a potent reward for loyal customers.

To investigate web users' participation in online communities and their consumption, we shall observe the web-based transaction factor as well. We shall also investigate the relationship between participation in online communities and consumption using transactional metrics. This could be the second contribution of this research. We perform customer segmentation by means of clustering methodology, one of the data mining tools. Through data mining methodology, which refers to extracting or mining knowledge from large amounts of data, we can search for interesting patterns in our dataset. We introduce the fundamental concept of "clustering" in the next chapter.

Table 3 E-Metrics proposed by Net Genesis

Terms	Explanation
Customer	A person who accesses a web site and purchases more than once.
User	A person who accesses a web site.
Click stream	The path of pages and links followed by a user during a visit to your web site.
Visit (Session)	A specific session at a web site that ends when the user has taken no further action after a given period of time, usually 30 minutes, indicating he or she is no longer "at" the site.
Page View	A request for a document (rather than an element such as an online image, movie, or audio file) on your web site.
Hit	A single entry in a server log file, generated when a user requests a resource on your web site. Requests are often referred to as "hits". A request can result in an error or a successful transmission of any data type.
Frequency	Frequency describes how often a prospect or customer performs a specific action.
Duration	Duration is a measure of time spent on a specific activity.
Visit Frequency	Scored based on number of visits per duration.
Visit Duration	Scored based on number of minutes per visit.
Purchase per visit	Scored based on amount of money spent per visit.

Table 4 Customer Segmentation by Duration (Song *et al.*, 2001)

User Type	Ratio	Frequency (6 months)	Average purchase per shopping (KRW)	Specifics
Heavy user	36.3%	4 times	60,000~70,000	Male, over 40
Medium user	37.9%	1.8 times	6,000~25,000	Female
Light user	25.8%	1.1 times	4,000~10,000	Below high school Novice in using Internet

Methodology

In this paper, we use cluster analysis for customer segmentation. Clusters can be used to group the subjects according to some measure of distance, relatedness, or similarity between importance ratings or utility measures (Aldenderfer and Blashfield, 1984). Once the clusters are identified, tests can be performed on various segmentation measures to identify the defining variables in the clusters (Green and Krieger, 1991). Punj and Stewart (1983) suggest two approaches for clustering. The first is hierarchical clustering, to determine the number of clusters, and the second is nonhierarchical clustering, for fine-tuning the results. The representative nonhierarchical clustering, the K-means algorithm with Euclidian distances, is superior to hierarchical clustering methods.

In comparison with nonhierarchical clustering, hierarchical clustering initially takes each observation and places it in its own cluster. Next, the two closest clusters are combined. This continues until there is only one cluster left. However, in most cases, an analysis is made to determine the appropriate number of clusters. Since hierarchical cluster analysis is extremely time consuming, it is rarely used in practice. Therefore, in this research, we use nonhierarchical clustering for the segmentation.

Clustering process

There are certain data cleaning and preparation procedures that should be performed prior to the clustering. We introduce a representative preliminary process for clustering.

Step1: Select the variables to use in the analysis

Used as a segmentation technique, cluster analysis should be performed on a small number of variables. The variable selection can be accomplished either by judgmental selection, or by a factor analysis. However, the selection process is still judgmental, so it is a good idea to select more variables than will be used in the final solution, then use cluster analysis to determine which ones work the best. After performing an analysis to determine

which variables to keep in the cluster analysis, there may still be several variables that need to be tested. The cluster means can be used to determine which of these variables is useful and which ones to drop, as well as to help with determining which ones are better than others when two variables are related to one another (Nargundkar and Olzer, 2000).

Step 2: Standardize the data

All variables included in the analysis must be standardized to a mean of 0 and a standard deviation of 1 because of the different units of variables. The reason for this is to put the variables on an equivalent scale. The weighted value is calculated for the analysis.

Step 3: Remove any data outliers

Standardizing the dataset makes this very easy. After the data is standardized, it remains only a matter of choosing how many standard deviations from the mean is too far, thus identifying outliers.

Step 4: Decide what type of cluster analysis to perform

Determine which cluster method is better. The nonhierarchical method is generally used in cases where there are a small number of observations (less than 100) and few variables.

Step 5: Decide number of clusters

While there is no perfect way to determine the number of clusters, there are some statistics that can be analyzed to help in the process (Milligan & Cooper, 1985). These are the Pseudo-F statistic, the Cubic Clustering Criterion (CCC), and the Approximate Overall R-Squared.

Step 6: Cluster validation

There is no right cluster analysis solution, just different viewpoints on the same dataset. Hence, there is no proven and universally accepted test that determines whether the produced solution is final or valid (Nargundkar and Olzer, 2000). However, in order to validate the cluster analysis, the following steps can be taken:

- Perform another cluster analysis using a different subset of variables and determine the optimal number of clusters.
- Compare the results from the above step to the findings achieved through the original cluster analysis with the original subset of variables.
- For data sets that are large enough, randomly split the data in half and perform cluster analysis on both sets of data, using the same subset of variables and number of clusters on the two halves, and compare the two solutions.

Empirical Analyses

Overall architecture of web log analysis

In general, we prepare the dataset used for this research with an overall architecture of web log analysis composed of three steps. The first step is gathering the necessary dataset. The necessary items of the access log file and registration data have been accumulated for a given research period. The second step is preprocessing, through which log data are cleansed and sessions (visits) are identified. Then, log data are matched with registration data. After searching each user's transaction data, the other specific data, such as user profiling and billing data, are combined with the corresponding session file. The last step is classifying this preprocessed data.

Data collection

To perform our empirical analysis, we collected web transaction data from an online pet shop. This shop specializes in pet supplies (e.g., food, accessories, cleaners, etc.), and maintains an online community by means such as a BBS (Bulletin Board System). Both member and nonmember users can purchase online goods either by credit card or by bank account transfer. We collected the web log data from March 22, 2002 to April 1, 2002, this constituting members' demographic information and users' online behavioral information.

The full dataset is analyzed in two stages. In the first stage, we integrate all the log access data by accessing time without distinguishing whether the user is a member or nonmember. In this analysis, we can be informed of the distribution of purchasing amount, duration per visit, and clicks per visit of both member and nonmember users. The full dataset contains 14,312 observations. Though not all the accesses in this dataset are significant, it nevertheless provides traffic data for the observed site.

Each variable has a different unit. For example, clicks are counted one by one when the user hits. Duration variable is measured in seconds, and is measured as the time difference between the first accessing time and the last accessing time. In the case of binary variables, a member user is described as "0" and a nonmember user as "1". The amount variable is in units of Korean Won. As the units of these variables differ from each other, we standardized the variables. The variable settings are shown in Table 5.

In the second stage, we use member data combined with their corresponding demographic data, which are available only from a registration record for each member, as shown in Table 6. At this site, users should fill in fields required for registration, such as resident registration number or address. We can find the sex and age of customers from the resident registration numbers. Other kinds of information about members such as job, education, and interests have been frequently omitted, which prohibits us from using them as variables. The address variable has no significance in this analysis, so it is automatically removed. The total number of objects in this dataset is 433, after the removal of outliers. Outliers may emerge as singletons or as small clusters far removed from the others. To do outlier detection, we remove clusters with less than ten objects.

In our dataset, three variables—region, access time, and job—are rejected automatically. The region variable has too many missing data, and the other two variables are not so important for this classification.

Data preparation

The dataset used in this study was obtained from an online shopping mall on an APACHE web server with a MySQL database. Every web page on this site is developed in PHP, which supports a session tracking function by checking user requests automatically when users access the web site. PHP is a widely used scripting language that is especially suited for web development, and can be embedded into HTML. This creation of user-defined functions enables us to make custom storage and retrieval handlers, making it possible to store the session data within any PHP-supported media, such as a MySQL database. Configuration of PHP's session-handling feature takes place in the php.ini file, among which the most important directives are as follows:

session.save_handler (files | mm | user).

In these directives, we can be informed of three methods for storing and retrieving session information; flat files (**files**), using shared memory (**mm**), and user-defined functions (**user**).

Originally, a session can be defined as the timeframe in which a user navigates the web site. PHP can track a user throughout the session by assigning a unique session identification number (SID) to that user. A default SID created by PHP looks like the following example:

fc94ad8b1ee49ef79c713ee98ac1fcc4

Table 5 Variable Settings of Full Dataset

	Name of Variable	Contents of Variable	Type of Variable
User Status	Member	Member/Nonmember	Binary
Click Stream Data	Click	Min 1 / Max 512	Interval
	Duration	Min 0 / Max 25,346 (seconds)	Interval
Purchase Amount	Amount	Min 0 / Max 119,000 (KRW)	Interval

Table 6 Variable Settings of Member Dataset

	Name of Variable	Contents of Variable	Type of Variable
Demographic Data	Member	Member/Nonmember	Binary
	Sex	Male/Female	Binary
	Age	Min 11 / Max 54	Interval
	Region	Seoul/Daejon/Pusan/Kwangju/Others	Rejected
Click Stream Data	Average Clicks per visit	Min 1 / Max 166	Interval
	Average Duration per visit	Min 0 / Max 4,528 (sec)	Interval
	Access per Period	Min 1 / Max 9 access per period	Interval
	Job	75% Missing	Rejected
	Access Time	0~6/ 6~12/12~18/18~24	Rejected
Purchase Amount	Amount	Min 0 / Max 143,850 KRW	Interval

From the timeframe data set, if an SID belongs to a member user, the customer profile data are merged with online behavioral data. In the case of a nonmember user, there is limited information apart from the number of clicks and duration per visit.

Preliminaries for cluster analysis

Variable selection

In our dataset, we did not perform an analysis to determine which variables to keep in the cluster analysis. Instead, variable selection is accomplished by judgmental selection because of limited web log data. However, variables from web transaction data such as access, duration, and click data are required for correlation testing. The cluster means of each standardized variable were similar, which suggested that we could select all the variables (Nargundkar and Olzer, 2000).

Number of clusters

For this analysis, we ran a separate analysis for each number of clusters. On the one hand, the number of clusters is close to the number of variables used. This is because with credit data, each variable included in the analysis will have a high value in one cluster. On the other hand, the statistics are offered to help determine the number of clusters; Pseudo-F statistics, the Cubic Clustering Criterion (CCC), and the Approximate Overall R-Squared. These are all measures of fit for the analysis (Nargundkar and Olzer, 2000). In general, by maximizing each value of these three statistics, the number of clusters is determined. In this paper, we choose the first method. The full dataset is divided into four clusters and the member dataset is divided into six clusters.

Results

Results of full dataset

In the analysis of the full dataset, the full dataset is divided into four clusters, as shown in Table 7.
Table 8 shows the results of a case cross-tabulation of the full dataset. This analysis shows clear characteristics for four clusters. First, as shown in cluster 3, most nonmember users tend to make short visits. This group shows a low

number of clicks and a short visit duration without a purchase. The ratio of this group to the full sample is 92.3%. Only four out of 13,203 observations purchase any goods, and the value of these is less than 20,000 KRW.

On the other hand, cluster 1 shows a high proportion of large purchase amounts with a duration of 10~30 minutes on the site. The ratio of members to nonmembers in this group is 6:4. We infer from this cluster that some online users buy goods without registration because of some reason such as the convenience of online shopping. Furthermore, both the duration and the number of clicks recorded for this cluster are below the average level, suggesting that the users of this cluster are only interested in buying without paying further attention to the site. Another interesting point is that some of them stay at this site for over an hour. We infer that nonmember users who are not accustomed to this site spend time searching for the goods they want.

Cluster 2 shows many nonmember users who have an interest in the site, judging from the duration figures. Unlike the users of cluster 3, the nonmember users in this group stay longer and their purchase ratio is higher than that of cluster 3. We infer that the users of this group have an interest in this site, but are not motivated to register yet. However, this group has a high potential to become loyal users.

Cluster 4, composed of more than 70% member users, shows that these users spend much time in this site without a significant purchase amount. We infer that the reason they spend more time than other users is that they have another purpose for visiting. We cannot specify the reason, but considering the characteristics of this site, several possibilities can be guessed. They may be participating in online communities or searching for better content.

As a result of previous research, we can see a clear classification of online users. Through the ratio of member to nonmember users, we can see that the member users have a greater willingness to pay. Some nonmember users in cluster 1 show high spending, but we cannot identify their IP addresses or their demographic information. However, based on the information of member users who visited this site during the research period, we could obtain a more detailed classification. We examine member users' results in the next section.

Results of member customer data set

As mentioned previously, the member data, combining members' demographic data with web transaction data, is divided into six clusters, as shown in Table 9.

In the table, the age variable is obtained from the registration number of member users. The sex variable is represented by "0" for female and "1" for male. The access variable indicates the numbers of member users visiting the site during the research period. The other variables are the same as for the analysis of the full dataset. All these variables are standardized as well.

In the member dataset, the number of observations is only 433, even though we find 830 members from the full dataset. The reason for the different number from the full dataset is that if the member customer visits repeatedly, this adds to the access count.

We can see the variable distribution of each cluster in Table 10.

All clusters are divided into six subgroups, as mentioned. Cluster 1 and cluster 3 are composed of women and cluster 2 is composed of men. Even though cluster 3 is also composed of women, customers in this cluster are older than those in cluster 1. The distribution of web transactions of this cluster is similar to that of the full group. The access data of both cluster 1 and cluster 3 show that more than 60% of customers visit once in 10 days, and that the maximum frequency is 4 in 10 days. Based on the duration data of both clusters, we infer that more than 50% of customers stay for less than 10 minutes per visit. The click data show a similar trend to those for duration in both clusters. The customers of cluster 1 spend more money at this site than the customers of cluster 3. Of the customers in cluster 3, 72% did not buy anything, unlike the 48% in cluster 1. We infer from this classification that the older

people show more buying power than the younger people, in spite of their similar visit frequencies and average durations per visit. The subgroup of male users has no specific characteristics compared to the full dataset.

Table 7 Descriptive Statistics of Full Dataset

	No.	Min.	Max.	Mean	Std. Deviation	Remarks
ZMember	14,312	0	1	.94	.23	0: member 1: nonmember
ZClick	14,312	1	1,012	8.60	29.54	Click Number
ZDuration	14,312	0	25,346	219.35	737.16	Second
ZAmount	14,312	0	119,000	366.28	3,801.07	KRW
Valid N (listwise)	14,312					

Table 8 Case Cross-Tabulation of Clusters in Full Dataset

		Cluster Number of Case Cross- Tabulation				
		1	2	3	4	Total
Status	Member	46	5	0	779	830
	Nonmember	29	79	13,203	171	13,482
Duration Per Visit (Second)	Below 600	10	27	11,813	297	12,147
	600 ~ 1,800	44	43	1,307	348	1,742
	1,800 ~ 3,600	14	6	83	192	295
	Above 3,600	7	8	0	113	128
Amount (KRW)	No Purchase	0	81	13,199	754	14,034
	0~20,000	0	2	4	190	196
	20,000~40,000	36	1	0	6	43
	Above 40,000	39	0	0	0	39
Clicks Per Visit	Below 10	15	0	11,656	303	11,974
	10 ~ 50	40	0	1,432	420	1,892
	50 ~ 100	16	0	101	165	282
	Above 100	20	84	118	247	469
Total		75	84	13,203	950	14,312

The customers of cluster 4 are characterized by frequent visits. They visit more frequently than other subgroups and stay for between 10 and 30 minutes per visit, on average. As for the purchasing amount of these customers, this cannot be defined succinctly. The data for this group are scattered over the intervals. We cannot judge from these data why they visit so frequently. However, the significant characteristic is the very low ratio of nonpurchase, suggesting that one of the important reasons for visiting is to buy something at this site. The age data of this group show a similar distribution to those of the full member dataset.

Table 9 Descriptive Statistics of Member Dataset

	No.	Min.	Max.	Mean	Std. Deviation	Remarks
ZAge	433	11	54	26.80	6.79	
ZSex	433	0	1	.17	.37	0 : Female 1 : Male
ZAccess	433	1	9	1.80	1.45	Visit Frequency
ZDuration	433	0	4,528	901.94	888.81	Seconds
ZClick	433	1	166	29.18	28.18	Click Number
ZAmount	433	0	119,000	8,595.38	16,039.79	KRW
Valid N (listwise)	433					

Table 10 Case Cross-Tabulation of Clusters and Variables in Member Data Set

		Cluster Number of Case						Total
		1	**2**	**3**	**4**	**5**	**6**	
Sex	Female	60		179	23	18	81	361
	Male		53		9	5	5	72
Age	10's	0	6	37	1	0	4	48
	20's~30's	12	20	142	26	18	66	284
	30's~40's	36	19	0	4	5	15	79
	40's~50's	12	8	0	1	0	1	22
Access	1	39	40	124	0	15	61	279
	2~3	19	12	50	0	5	21	107
	4~6	2	1	5	21	3	4	36
	Above 7	0	0	0	11	0	0	11
Duration (Second)	0~600	35	29	118	6	15	2	205
	600~1,800	24	20	59	23	8	40	174
	1,800~3,600	1	3	2	3	0	30	39
	Above 3,600	0	1	0	0	0	14	15
Amount (KRW)	No Purchase	29	29	130	5	0	27	220
	Below 20,000	25	18	46	17	0	50	156
	20,000~ 40,000	5	5	3	6	5	8	32
	Above 40,000	1	1	0	4	18	1	25
Clicks	Below 10	21	20	75	4	11	0	131
	10 ~ 50	38	27	100	22	11	24	222
	50 ~ 100	1	6	4	6	1	46	64
	Above 100	0	0	0	0	0	16	16
Total		60	53	179	32	23	86	433

We infer, on the basis of the statistics of this cluster, that the purchasing amount has a positive relation to the frequency of visits. The more the member users visit, the more chance to purchase they have; and they purchase more than average compared to other clusters, which means that the users in this cluster are loyal users. Their average amount of purchase per visit is around 3,000 KRW. This means that the users in this cluster purchase goods at this site whenever they visit. Although this is low compared to cluster 5, the high-spender group, the users in this group are more valuable for the site.

The customers in cluster 5 are characterized by the highest ratio of purchasing amount, and are thus regarded as the most loyal customers in the view of the company. They purchased at least 20,000 KRW on average. Web transaction data of this cluster show a similar distribution to those of the full dataset. The access data of this group show that 50% of the customers visit once and that the maximum number of visits is five. It can be inferred that customers whose only purpose is online purchasing do not always visit so frequently. The age data of this group are focused around 20~40.

We infer, on the basis of the statistics for this cluster, that the high spenders neither visit frequently nor stay long. Their purpose for visiting is clear, and they are not interested in other activities or contents at this site. We cannot identify the reason for purchasing at this site; it is important to discover this reason, by surveying, and to offer more motivation for buying.

The customers in cluster 6 are characterized by long duration. They stay more than 10 minutes at least. The ratio of the nonpurchasing group is 31%, which shows they have another reason for visiting this site, such as interesting contents or virtual community. On the other hand, more than 50% of this group buys less than 20,000 KRW worth of goods at this site. The age distribution in this group is similar to that of the full member dataset. There is no one who clicks less than 10 times per visit in this cluster.

We infer, on the basis of the statistics of this cluster, that the users in this cluster have interests in this site other than purchasing. They do not visit frequently, so we cannot say that they participate in activities at this site. In addition to this, we infer that their purchasing propensity is on the low side, or that they are prudent in this respect. The users in this cluster purchase their goods after searching the site thoroughly. Based on above segmentation analysis, we can divide this dataset into several meaningful subgroups, as shown in Table 11.

Conclusion

The relationships between customer behavior and revenue were studied by statistical analysis of log data of a sample online shop in Korea. In doing this, we employed clustering analysis.

In the analysis of the full data set, including member and nonmember users' behavioral data, we found empirical evidence that only 5.7% of the dataset represents members. We cannot count unique visitors as nonmember users, because we are not able to recognize nonmembers' IP addresses precisely. However, we can gain a rough idea of nonmembers' reach ratio for this site.

For more detailed analysis of customer behavior at this site, we have examined the online transaction data of member users. We have 433 objects in this dataset in total because member users' repeated visits have been accumulated in the access data. Through this number, we can infer the number of unique users among the member users of this site. In addition, we can analyze this with the demographic dataset.

To summarize, the main findings of the study are as follows:

1. In the full dataset, all the data are divided into four groups. The click-to-visit ratio of this site is very low, but some nonmember users spend more than average. This means that in online retail shops offering online communities, customers do not always visit and purchase as members who participate in an online community.
2. In the member dataset, clusters 1, 2, and 3 are segmented by demographic variables. There are no specific relationships between demographic and transaction variables in these clusters. However, 80% of member users are female and divided into two groups by age. The users in these clusters have different needs regardless of transaction. Therefore, online retailers should have different marketing strategies for them.
3. In the member dataset, clusters 4, 5, and 6 are segmented by transaction variables. Each cluster shows clear differences in visit frequency, purchasing amount, and visit duration. We can infer that visit frequency affects purchasing amount to some extent. Therefore, the users in cluster 4 have the potential to become loyal users.
4. The users in cluster 5 purchase more than others. These customers do not always stay longer than others. This means that they visit for purchases but seldom participate in the community. We infer that they are profitable but not regular. Therefore, retailers should try to fortify site loyalty by providing content that might interest this group.
5. Of the users in cluster 6, who stay longer than others, one in three does not purchase at this site. We can infer that they have other reasons for visiting, or that they are very prudent consumers. In spite of their average amount of purchase, they are interested in the site. That is, they could be very sensitive to marketing promotions or strategic selling methods. Marketing event focusing is needed for this group, such as organizing collective purchases for special interest groups.

Limitations of study

In this paper, we could not include members' employment or education variables, because of missing raw data. This prevents us from doing more detailed customer segmentation. Furthermore, we collected only 10 days of web log data owing to hardware problems with the site. We leave more detailed and precise customer segmentation to a subsequent study.

Table 11 Customer Segmentation of Sample Online Pet Shop

Cluster	% of Sample	Nickname	Description
1	14%	Older women	Mainly composed of women in their 30sVisits 1.4 per 10 days Stays 9 min. on average. Spends 6,800 KRW on average Average age 35.2
2	12%	Men's Group	Mainly composed of men with normal distribution of age Visits 1.3 per 10 days Stays 12 min. on average Spends 6,4000 KRW on average Average age 30
3	41%	Younger Women	Mainly composed of young women Visits 1.4 per 10 days Stays 8.5 min. on average Spends 2,700 KRW on average Average age 22
4	7%	Frequent User	Visits frequently with medium purchasing amount Visits 5.8 per 10 daystars 15 min. on average Spends 17,000 KRW on average Average age 26
5	5%	High Spender	Visits 1.8 per 10 daystars 10 min. on average Spends 58,000 KRW on average Average age 27
6	19%	Long Stay	Visits 1.5 per 10 daystars 35 min. on average Spends 7,200 KRW on average Average age 26

References

1. Aldenderfer, M.S. and Blashfield, R.K. (1984), *Cluster Analysis*, Sage Publications, Newbury Park, CA.
2. Büchner, A.G. and Mulvenna, M.D. (1998), "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining," *SIGMOD Record* 27(4): 54–61
3. Cooley, R., Mobasher, B. and Srivastava, J. (1999), "Data Preparation for Mining World Wide Web: Browsing Patterns," *Journal of Knowledge and Information Systems*, 1(1): 5-32
4. Donthu, N. (1999), "The Internet Shopper," *Journal of Advertising Research*, 39(3): 52-58
5. Emmanouilides, C. and Hammond, K. (2000), "Internet Usage: Predictors of Active Users and Frequency of Use," *Journal of Interactive Marketing*, 14(2) : 17-32
6. Englis, B.G. and Solomon, M.R. (2000), "Life/Style OnLine©: a Web-Based Methodology for Visually-Oriented Consumer Research," *Journal of Interactive Marketing*,, 14(1): 2-14
7. Fader, Peter S. (2001), "Web Metrics: Making the Most of Your E-commerce Data," *Journal of Interactive Marketing*, 15(1)
8. Forrester Research (2000), *Three Types of Retailers and Their Recommended CRM Strategies*, TechStragegy Report
9. Green, P.E. and Krieger, A.M. (1991), "Segmenting markets with conjoint analysis," *Journal of Marketing*, 55(4): 20–31
10. Hanson, W. (2000), *Principles of Internet Marketing*, Australia ; Cincinnati, Ohio : South-Western College Pub., 2000
11. Hao, M. C., Ladisch, J., Dayal, U., Hsu, M. and Krug, A. (2001), "*Visual Mining of E-Customer Behavior Using Pixel Bar Charts*," HP Labs Technical Reports
12. Hoffman, D.L. and Novak, T.P. (1999), "Measuring the Customer Experience in Online Environments: A Structural Modeling Approach," *Marketing Science*, 19(1): 22–44
13. Kozinets, R.V. (1999), "Tribalized Marketing?: The Strategic Implications of Virtual Communities of Consumption," *European Management Journal*, 17(3): 252–264
14. Leeflang, P.S.H. and Wittink, D.R. (2000), "Building Models for Marketing Decisions: Past, Present and Future," *International Journal of Research in Marketing*, 17(2-3): 105–126
15. Levin, N.(2001), "Predictive Modeling Using Segmentation," *Journal of Interactive Marketing*, 15(2): 2-22
16. Mahajan, V. (2000), "Marketing Modeling for E-business," *International Journal of Research in Marketing*, 17(2-3): 215–225
17. Milligan, G.W. and Cooper, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, 50(2): 159–179
18. W. Moe, (2001), "Capturing Evolving Visit Behavior in Clickstream Data," Working Paper, The Wharton School
19. Nargundkar S. and Olzer, T.J. (2000), "An Application of Cluster Analysis in the Financial Services Industry," Case Study, Sigma University
20. Punj, G. and Stewart, D.W. (1983), "Cluster Analysis in Marketing Research: Review and Suggestions for

Application," *Journal of Marketing Research*, 20(May): 134–148

21. Rosenstein, M (2000), "What is Actually Taking Place on Web Sites: E-Commerce Lessons from Web Server Logs," *ACM Conference on Electronic Commerce* (EC-00), October 17-20, 2000, Minneapolis, Minnesota, USA
22. Rothaermel, F.T. and Sugiyama, S. (2001), "Virtual Internet Communities and Commercial Success: Individual and Community-Level Theory Grounded in the Atypical Case of TimeZone.com," *Journal of Management*, 27(3): 297–312
23. Song, H.S, Kim, J.K. and Kim, S.H. (2001), "Mining the Change of Customer Behavior in an Internet shopping Mall," *Journal of Information Management*, 10(1): 157-168
24. Weinstein, A. (2000), "Segmentation: Developing Target Markets," *Handbook of Online Marketing Research*, Grossnickle, J. and Raskin, O., New York: McGraw-Hill, 2001 \

Notes