Review Article

Getting Business Insights through Clustering Online Behaviors

Jounghae Bang,¹ Yoonho Cho,¹ and Min Sun Kim²

¹College of Business Administration, Kookmin University, Jeongneung-ro 77, Seongbuk-gu, Seoul 136-702, Republic of Korea

²School of Tourism and Distribution Management, Hyupsung University, Sang-ri, Bongdam, Hwasung City, Kyunggi-do 445-745, Republic of Korea

Correspondence should be addressed to Min Sun Kim; sunnyminkim@hanmail.net

Received 31 August 2014; Accepted 16 November 2014

Academic Editor: James J. Park

Copyright © 2015 Jounghae Bang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This study aimed to explore the online users' behaviors. Since the Internet was introduced to the market, the various and frequent online activities have increased, and it becomes more important for the businesses to understand the online users. Therefore this study analyzed the online users' behaviors and segmented the users by using *K*-means clustering method using actual clickstream data. There were four different research questions and, thus, four different sets of segmentations. It was found that many people find much of entertaining from online using SNS, games, and so on. In addition, some people only have access to a few specific websites. Some use the online service regularly every day while others use it in a very irregular pattern. People were divided into two groups, weekday group and weekend group. People are likely to be using the Internet either on weekdays or at weekend. Teenagers and people in their 50s are more likely to use it during weekend. In addition, teenagers also show different time zone (e.g., overnight) to use the Internet from other age groups. These results can shed light on understanding what consumers do online and what they are interested in currently and on decision making in marketing strategy.

1. Introduction and Research Questions

As the Internet technology has been advanced, the usages of Internet service have been rapidly increasing. People in the USA have used the online service longer ever since the Internet introduced. It increased from 11.9 hours to 12.9 hours per month. In addition, people use online service in many different areas such as download, games, and shopping [1–3]. As the use of the Internet increases, the users have become more active online. Therefore, it has become important for the businesses to understand the online users.

There have been many studies done on online behavior. Most of the studies on the Internet and the users have tried to understand who they are and what they are doing online [4–6]. These studies conducted segmentation and grouped the online users by such variables as age, gender, and lifestyle [7–9]. The studies have used mostly survey data, which indicates what online users perceived they did online.

This study, however, used actual clickstream data, which can give us the clear ideas about what they have done instead of what they perceived they have done. With the clickstream

data, proper variables were identified. Using the *K*-means clustering method, this study attempts to find different groups of online users so that customized services and value can be provided to the users based on their patterns and preferences of websites and their functions.

We attempted to answer the following research questions.

- (Q1) What do the online users use the Internet for?
- (Q2) What are the characteristics of online behaviors?
- (Q3) When do the online users use mostly the Internet during the week and weekend?
- (Q4) When during a day do the online users usually use the Internet?

2. Experiment and Results

2.1. Data Description. The data analyzed was from 5,000 online users, and about 150 millions of transactions from July 2012 to June 2013 were used. The transactions include user id, time and date, accessed URL, and dwell time.

2.2. Analysis Method. Clustering is one of the data mining methods that divide multivariate data into several different groups by its similarities of characteristics. Clustering is widely used to understand the characteristics of consumers and markets in the business domain, and algorithms used are *K*-means clustering, Kohonen Network, TwoStep Clustering, and so on. As other studies [5, 6, 10], this study also used the *K*-means clustering method with IBM SPSS modeler 15.0 to segment the markets. The *K*-means clustering method is relatively easy to analyze and effective to use different forms of data while it is difficult to decide the appropriate number of clusters. In order to choose the optimal numbers of clusters, we used Silhouette measure which is the average, over all records $(B-A)/\max(A, B)$, where A is the record's distance to its cluster center and B is the record's distance to the nearest cluster center that it does not belong to.

According to the four research questions above, four different sets of segmentations were found. They are (1) topic segmentation, (2) behavior characteristics segmentation, (3) weekday segmentation, and (4) time segmentation.

2.3. Results. (Q1) What do the online users use the Internet for? Topic segmentation.

To answer the first question, data was clustered by the topics of websites that users visited. Table 1 shows the result of topic segmentation, which found five different segments.

The largest group is Cluster-4, "information consuming and entertainment group," which includes 38.7% of sample, 1,933 persons. They are mostly the main users of the Internet, who use search, community, SNS, entertainment, game, other services, and business and economics related websites.

Cluster-1 is "portal service group." They mostly use the portal services in the Internet, which account for 19.3% of total sample. "Finance and public service group" is Cluster-2. They use mostly finance information and public services, and this group is the smallest with 3.8% of total sample (191 persons). Cluster-3 is "email and news group" with 18.6% of total sample, and they are users of email, news, and Internet and computer related websites. Cluster-5 is "shopping group," of which people use the Internet mostly for shopping and they account for 19.7% of total sample. Therefore, around 20% of online users enjoy mostly online shopping, not other Internet services.

(Q2) What are the characteristics of online behaviors? Behavior characteristics segmentation.

Second set of segmentation is about the behavioral patterns of online users. The variables for this clustering are COV, COVERAGE, D_COV, SCH_KEYWORDS, and VDAYS. The explanation of each variable is provided in Table 2. With those variables, four segments were found as shown in Table 3.

The largest group in Table 3 is Cluster-4, "various daily access group." It includes 46.8% of total sample. In this group, variation in daily Internet usage is big. That is, the online users are not using the Internet consistently during a year. For some days, people use it a lot, while in some other days they do not use it much. Since it is the largest group, mostly women and men belong to this "various daily access group" (46% of men and 48% of women).

People in Cluster-1 are in "everyday access group," the second largest group (around 20% of total sample). They access the Internet everyday consistently often, but not many different websites.

People in Cluster-2 are in "a few important websites access group," the third largest group with 17.5% of total sample. They use only small number of websites and they tend to visit those websites in a desultory fashion. They only use the Internet for a few topics.

The last group of people in Cluster-3 is in "many different websites access group." They are likely to search many different keywords, and thus they are most likely to visit many different categories of websites. It is the smallest group with 15.8% of total sample.

For "various daily access group" and "many different websites access group," higher percentage of women are included than percentage of men. On the other hand, "a few important websites access group" and "everyday access group" have higher percentage of men than percentage of women.

With all ages, higher percentages of people are included in "various daily access group." In particular, for teenagers, 68.9% of them are in this group, while for other ages mostly 44~54% of them are in the group. For people aged 10s and 20s, "many different websites access group" is the second largest group with 11.5% of people aged 10s and 23.3% of people aged 20s. However, for people aged 30~50s, the second largest group is "everyday access group" and then "a few important websites access group."

(Q3) When do the online users mostly use the Internet during the week and weekend? Weekday segmentation.

To answer the question, the Internet usage in each day during week is calculated. As shown in Table 4, two groups were found. Cluster-1 is "weekdays group" and Cluster-2 is "weekend group." As expected from the title of group, "weekdays group" uses the Internet usually during weekdays from Monday to Friday, while "weekend group" uses the Internet during weekend from Saturday to Sunday more often.

People in their 10s and 50s are using the Internet during weekend slightly more than weekdays.

In "weekdays group" around 60% of people aged 20s are included. For people aged 20s, 30s, and 40s, more users are using the Internet during weekdays (more than 50%). However, as age increases, percentages in "weekdays group" are decreasing while more people use the Internet at weekend. And finally for people in their 50s, "weekend group" (51.3%) is slightly higher than "weekdays group" (48.7%).

(Q4) When during a day do the online users usually use the Internet? Time segmentation.

Three groups were found as shown in Table 5. A group of people who are in Cluster-1 are using the Internet usually afternoon time, from noon to 11 pm which is called "afternoon group." However they are usually not using the Internet much. Cluster-2 is "overnight group" which is using the Internet from 6 pm to 5 am. Cluster-3 is "working hour group" who uses the Internet from 6 am to 5 pm.

"Afternoon group" have shown that they are not the heavy users, which account for 31%. Almost half of people (45%) use the Internet during working hours, from 6 am to 5 pm. The smallest group of people is "overnight group." They stay up all night using Internet.

TABLE 1: Topic segmentation.

Clusters					
Input (predictor) importance

Cluster	1	2	3	4	5
Label	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster-5
Description	Portal service group	Finance and public service group	Email and news group	Information consuming and entertainment group	Shopping group
Size
Inputs

TABLE 1: Continued.

TABLE 2: Segmentation variables for behavior characteristics.

Variable	Explanation
COV	The coefficient of variation of dwell time per site (“standard deviation/average” of dwell time per site)
COVERAGE	The degree to which each individual visits diverse category websites among the 1231 website categories
D_COV	The coefficient of variation of dwell time per day (“standard deviation/average” of dwell time per day)
SCH_KEYWORDS	The number of different keywords that each individual searches in a year
VDAYS	Total number of days to access the Internet in a year

TABLE 3: Behavior characteristics segmentation.

Clusters				
Input (predictor) importance

Cluster	1	2	3	4
Label	Cluster-1	Cluster-2	Cluster-3	Cluster-4
Description	Everyday access group	A few important websites access group	Many different websites access group	Various daily access group
Size	19.9% (997)	17.5% (873)	15.8% (791)	46.8% (2339)
Inputs

As expected, 62% of teenagers are in “overnight group,” while all other ages (20s~50s) are mostly (around 50%) in “working hours group.” 32% of men are in group 1 and 25.2% of men are in group 2, while 29.9% of women are in group 1 and 22.2% of women are in group 2.

For “working hours group,” 47.9% among women are included, which is higher percentage than men (42.7% among men are included in “working hours group”). However “overnight group” and “afternoon group” show higher percentages among men than the percentage among women.

3. Implication and Concluding Remarks

This research aims to identify better segmentation of online users based on their behaviors using actual clickstream data. In order to understand the online users’ behaviors better, four different research questions were asked.

First, what do the online users use the Internet for? The results indicate that around 40% of people are enjoying entertaining, having relationships with people through online, information search, and business related websites.

TABLE 4: Weekday segmentation.

Clusters		
Input (predictor) importance		
	1.0	0.4
	0.8	0.2
	0.6	0.0
Cluster	1	2
Label	Cluster-1	Cluster-2
Description	Weekdays group	Weekend group
Size	54.3% (2717)	45.7% (2283)
Inputs	DF_Sat	DF_Sat
	DF_Sun	DF_Sun
	DF_Thu	DF_Thu
	DF_Tue	DF_Tue
	DF_Fri	DF_Fri
	DF_Wed	DF_Wed
	DF_Mon	DF_Mon

That is, 40% of people use the Internet for various general uses. On the other hand, three other groups were found who are narrowly focused on certain use of the Internet such as a shopping focused group, a portal focused group, and a computer, email, and news focused group. In addition small portion of people (3.8%) were found to use the Internet for finance related and public services. In sum, many people who use the Internet enjoy the various entertainment elements while some are focused on specific things. Therefore, it is considerable to develop new contents for people with

different interests based on what they are interested in while being online and what they usually do online.

Second, what are the characteristics of online behaviors? People were divided into two groups in big category, which are a group with various usages and the other group with a few topics focused. Most of the people, around 80%, are found to be a few topics focused. They are further divided into three groups; “various daily access group,” “everyday access group,” and “a few important websites access group”. On the other hand, a group with various usages is called

TABLE 5: Time segmentation.

Clusters			
		Input (predictor) importance	
		1.0	0.4
		0.8	0.2
		0.6	0.0
Cluster	1	2	3
Label	Cluster-1	Cluster-2	Cluster-3
Description	Afternoon group	Overnight group	Working hour group
Size	31.3% (1565)	24.0% (1201)	44.7% (2234)
Inputs	HF0005	HF0005	HF0005
	HF0611	HF0611	HF0611
	HF1217	HF1217	HF1217
	HF1823	HF1823	HF1823

“many different websites access group” and they are likely to search many different keywords, and thus they are most likely to visit many different categories of websites. Based on this different patterns of behaviors, the decisions about how to communicate with consumers via online can be made.

Third, when do the online users mostly use the Internet during the week and weekend? There were two groups found, which are “weekdays group” and “weekend group.” Teenagers and people in their 50s are included in “weekend group” slightly more than in “weekdays group” while people in their 20s~40s are more in “weekdays group” than in “weekend group.” Among the people in their 20s~40s, people aged in 20s are likely to be in “weekdays group” more than people aged in 40s. It indicates that as people become older, they do not have enough time for the Internet during weekdays. It implicates that for teenagers it is good to have promotion during weekend. During weekdays office workers can be a target for communication or promotions.

Fourth, when during a day do the online users usually use the Internet? Almost half of people (45%) use the Internet during working hours, from 6 am to 5 pm. 62% of teenagers use the Internet overnight. It implicates that, for different age groups, the different time zones for on-time communication would be worth considering.

In sum, the study attempted to explore the online users’ behaviors now that it has been more than 20 years since the Internet was introduced to the market using the clickstream

data which is not the perceptions of users. It may shed light on understanding online behaviors of people and help the businesses make better strategy to communicate with their customers using the online devices.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

- [1] H. Elshaafi and D. Botvich, “Trustworthiness inference of multi-tenant component services in service composition,” *Journal of Convergence*, vol. 4, no. 1, pp. 31–37, 2013.
- [2] I. H. M. Hwon, “Mining consumer attitude and behavior,” *Journal of Convergence*, vol. 4, no. 2, pp. 29–35, 2013.
- [3] The McKinsey Quarterly, “What do broadband consumers want?” *Web Exclusive*, vol. 4, 2001.
- [4] “Categorizing Online Shoppers,” November 2000, <http://www.emarketer.com>.
- [5] J. H. Lee and M. K. Choi, “Repurchase intention of experienced buyers in the internet shopping mall by using customer segmentation,” *Journal of Information Technology Applications & Management*, vol. 10, no. 1, pp. 19–34, 2003.
- [6] S. Y. Park and J. Y. Lee, “The analysis of on-line product categories based on consumer segmentation characteristics,”

Journal of Korean Distribution and Management, vol. 10, no. 1, pp. 59–84, 2005.

- [7] J. Arshad, P. Townend, and J. Xu, “A novel intrusion severity analysis approach for Clouds,” *Future Generation Computer Systems*, vol. 29, no. 1, pp. 416–428, 2013.
- [8] P. Kotler and G. Armstrong, *Principles of Marketing*, Prentice Hall, Englewood Cliffs, NJ, USA, 1989.
- [9] K. R. Kim and N. M. Moon, “Content modeling based on social network community activity,” *Journal of Information Processing System*, vol. 10, no. 2, pp. 271–282, 2014.
- [10] B. Loganayagi and S. Sujatha, “Enhanced cloud security by combining virtualization and policy monitoring techniques,” *Procedia Engineering*, vol. 30, pp. 654–661, 2012.