論文

Clustering of Public Library Users by Similarity of Visiting Paths Using Location Information

Noriko Sugie

[Abstract] This study aims to classify library users' behavior by analyzing location information acquired using observation methods based on the radio-frequency identification (RFID) technology. In 2012, I conducted a study at the Chiyoda Public Library, Japan, to track user visiting paths using an RFID system and administered a questionnaire survey to acquire user attribute data. The following data were obtained: time spent in the library, time spent browsing in each library zone, time spent browsing books by subject, visited points, and visiting paths. The data regarding the users' visiting paths were analyzed, after which the questionnaire responses were analyzed for each group. Hierarchical clustering based on the similarity among the character strings generated from the users' visiting paths identified two clusters. The users in Cluster 1 were those who would likely look for materials to borrow without sitting and therefore would visit more points with RFID tags. These users visited the general book zones, in which most materials available for lending are located. Conversely, the users in Cluster 2 were those who would likely look for materials and then sit down to read them. Therefore, they visited fewer points with RFID tags than the users in Cluster 1.

[Keywords] public library, user study, clustering, information-seeking behavior.

1. INTRODUCTION

Many studies have investigated information-seeking behavior in the context of information behavior research (Julien, 2000 & 2011). However, only a few studies have attempted to understand user behavior itself in a physical library, which is critical for understanding a user's information-seeking behavior in a library. This inadequate research is likely because of the complexity of user behaviors and the difficulty of collecting and analyzing user behavior data. Information on user behavior is indispensable to develop an understanding of library users and to improve library services.

2. LITERATURE REVIEW

Over the last decade, extensive research on customer behavior patterns has been conducted in the marketing field using radio-frequency identification (RFID) systems (Sorensen, 2003; Larson, Bradlow, & Fader, 2005; Hui, Fader, & Bradlow, 2009). The researchers installed RFID tags in retail stores or on shopping carts, collected customers' traveling data, and analyzed the shopping process. These studies showed that the RFID system is able to collect large volumes of accurate data such as customer location, time spent in a retail store, and travel path. Since customer behavior in a retail store is similar to user behavior in a library, it is considered that the methods of these studies can be applied to a study in a library.

Therefore, in 2010, I conducted an experiment in which an RIFD system was used to track users' visiting paths at a university library. In 2012, I conducted an extended study that improved the equipment and method used to collect data at a public library (Sugie, 2013; Sugie, 2012). The present study performs a statistical analysis of the quantitative data acquired in 2012 to discover the information-seeking patterns of users.

3. METHOD

This study aims to classify library users' behavior by analyzing location information acquired using RFID-based observation methods. Here, the term "information-seeking behavior" is defined broadly as behavior through which the user looks for information, including browsing and reading information resources in the library.

3.1. Data collection

This study was conducted on the 9th floor of the Chiyoda Public Library in Tokyo, Japan from April 2012 to May 2012. Users who agreed to participate in the study were given an antenna for receiving the radio waves emitted from the tags and a personal digital assistant (PDA) to record the data (Figure 1 and Figure 2), after which they proceeded to use the library as usual (behavioral investigation). The 9th floor contains 120,000 books and magazines, each of which has an RFID tag (Figure 3 and Figure 4). Alphabetic character in Figure 4 correspond to the character which means library zones in Table 1. These tags are used for security purposes in the library, but a nearby antenna can pick up the radio waves from the tags, making it possible to identify a user' s location in the library.

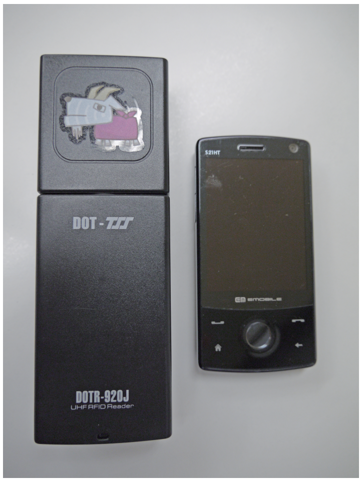

Figure 1. Antenna (left) & PDA (right).

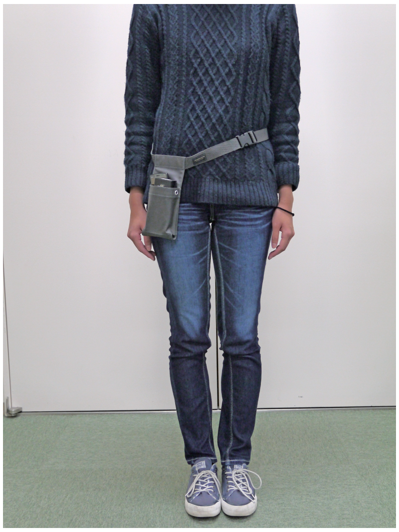

Figure 2. A user with Antenna & PDA.

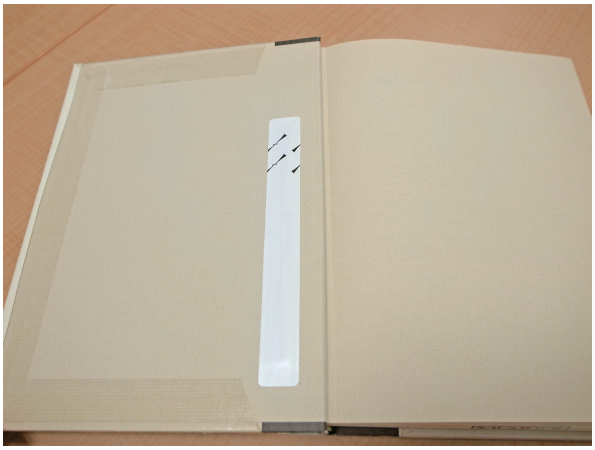

Figure 3. RFID tag (953MHz) on a book.

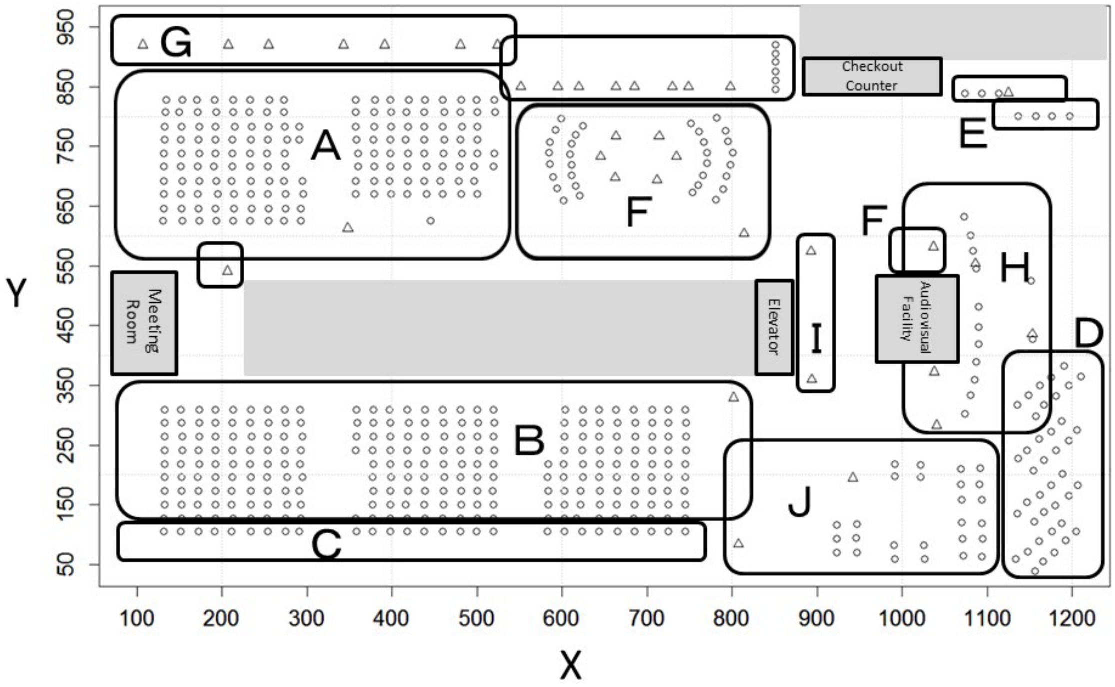

Figure 4. Locations of RFID tags and zones.

Table 1. Library zones.

Character	Zone
A	Research zone
B	General book zone
C	General book zone
D	Paperback, new books
E	Books returned
F	Information search zone
G	Reading chairs
H	Space for display
I	Library entrance
J	Magazine zone

Thus, when users travel with an antenna, time-series data about their location can be gathered. Thirty-four RFID tags were installed individually at locations where the radio waves of the RFID tags on the books were not able to reach (radio waves reach a few meters in any direction). The RFID system chronologically provides the time and ID number of the RFID tags at the points visited by users. By reffering each ID number recorded on a tag to the bibliographic data of the library, the shelf number, position coordinates on the floor map, and zone in the library were derived (Table 2).

A brief questionnaire survey on library usage was conducted immediately after users finished using the library. The questionnaire items covered user attributes (age, sex, and occupation), visit frequency, purpose of the visit, whether they borrow materials, and whether they sit on a chair or a sofa.

Table 2. Examples of data collected for a user.

Receipt Time of Tag Signals	ID Number of Resources	Shelf Number	Call Number	Title of Books	X-cordinate	Y-cordinate	Zone	Kinds of Sources
2012/05/01 18:45:08	130349186	22A-8	914.6	猫の一年	725	196	general book	primary sources
2012/05/01 18:45:09	1000039311	23B-2	914.6	日本人の愛したことば	724	173	general book	primary sources
2012/05/01 18:45:09	130349186	22A-8	914.6	猫の一年	725	196	general book	primary sources
2012/05/01 18:45:10	1000017754	21A-8	913.6	太宰治賞	746	242	general book	primary sources

3.2. Data analysis

3.2.1. Analysis process

Of the data obtained, the position coordinates of the tags were used to identify user groups by clustering the users’ visiting paths by similarity. The position coordinates of each user were converted into alphabetic characters referring to zones, as the position coordinates could not be analyzed in their original form. This process generated character strings for each user, each of which describes the user’s visiting path as a series of zones. For example, the character string “FFFFGGA” indicates that the user sequentially visited the information search zone, reading chairs, and research zone. In addition, radio waves were received from the RFID tag four times for zone F, twice for zone G, and once for zone A. These frequencies at which the antenna received radio waves from tags are regarded as the visiting frequencies for each point where RIFD tags are located in the library.

The edit distances were calculated from the character strings to express the degree of similarity among the visiting paths of users. The calculation of the edit distance is described in the next section. Clustering the users’ visiting paths using Ward’s method was conducted to identify user groups. The features of each group identified via clustering were analyzed with reference to the questionnaire responses. R version 3.02 was used for statistical processing.

3.2.2. Edit distance calculation

The edit distance is the minimum number of times a character string must be edited to convert it into another character string. This metric describes the dissimilarity between two character strings. One of the most basic edit distances is the Levenshtein distance. The Levenshtein distance is calculated from three editing operations: character insertion, character elimination, or replacement of a character with another character in a character string. For example, the Levenshtein distance between “ABC” and “BBDC” is two, i.e., A is converted into B, and D is inserted between B and C. In this study, the edit distances were normalized to the total number of characters in the converted character strings, because this number differs for each user.

4. RESULTS AND DISCUSSION

4.1. Research subjects

The library visit frequencies and occupations of the research subjects are shown in Table 3. The percentage of users who visited more than once a month was 74.6%. Of these users, the percentage of office workers who visited more than once was 82.1%, the highest for any occupation.

Table 3. Library-visiting frequency and occupation.

Occupation	More than once a month		Other		Sum	
	Num.	%	Num.	%	Num.	%
Office worker	78	82.1	17	17.9	95	45.5
Student	28	65.1	15	34.9	43	20.6
Unemployed	14	63.6	8	36.4	22	10.5
Home-maker	10	76.9	3	23.1	13	6.2
Others	26	72.2	10	27.8	36	17.2
Sum	156	74.6	53	25.4	209	100

4.2. Staying time

The first and last times that a user's antenna received radio waves were obtained for each user via behavioral investigation. These times were treated as the entrance and exit time, respectively, and were used to calculate the staying time for each user. The total staying time for all users was 681,354 seconds (189 hours, 15 minutes, and 54 seconds), and the average staying time was approximately 3,260 seconds (54 minutes, 20 seconds). Table 4 shows these times and the fundamental statistics.

Table 4. Fundamental statistics for staying time.

Number of users	Total time (sec)	Fundamental statistic				
		Min.	Median	Mean	Max.	SSD
209	681,354	465	2,459	3,260.1	16,190	2499.64

4.3. Groups generated via visiting path similarity

4.3.1. Clustering based on edit distances between users

A distance matrix comprising 21,736 edit distances calculated for 209 users in all combinations was generated. Ward's hierarchical clustering was conducted on the basis of the edit distances or degree of dissimilarity of the users' visiting paths, and a dendrogram was obtained (Figure 5). On the basis of the distances between the generated clusters, the dendrogram was divided by a length of 12, yielding two clusters. Cluster 1 included 151 users, while Cluster 2 included 58 users. The data regarding the users' visiting paths were analyzed, after which the questionnaire responses were analyzed for each group.

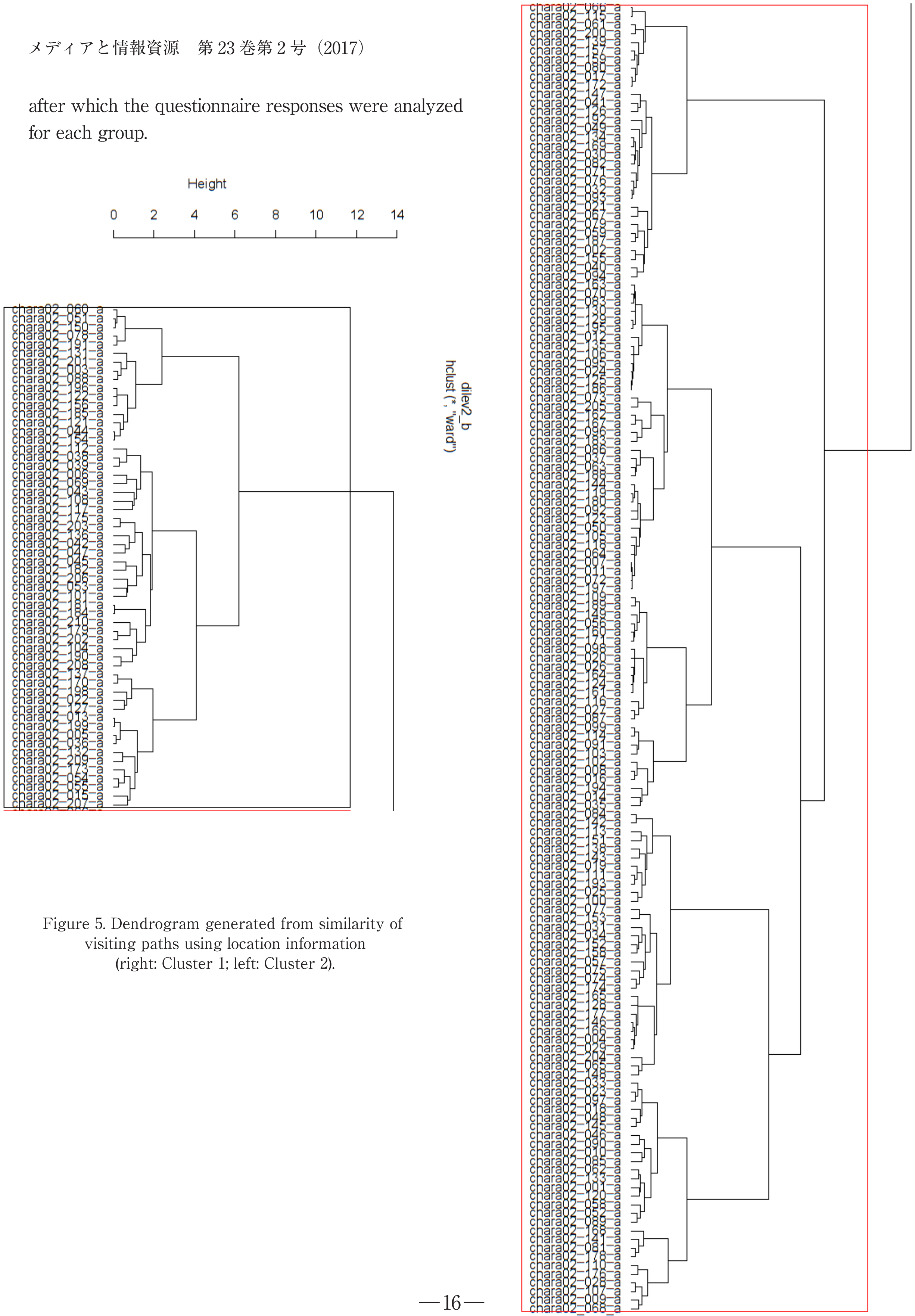

Figure 5. Dendrogram generated from similarity of visiting paths using location information (right: Cluster 1; left: Cluster 2).

4.3.2. Analysis of user visiting path data

The edit distance value directly affects the cluster construction. As stated above, the edit distance is determined by the number, type, and order of characters. Therefore, I calculated the sum of the number of characters and the type of characters for each group and analyzed the features of users' visiting paths that were reflected in the construction of these clusters. The order of the characters in the three elements that reflect cluster construction was not considered in this paper because the sum of the order of characters cannot be obtained.

(A) Frequency of the visited points and staying time

The fundamental statistics of the frequencies at which the users' antennas received radio waves from the RFID tags (visit frequency) were calculated for each cluster (Table 5). Both the mean and median values of the total visit frequency in the library for Cluster 1 were higher than those for Cluster 2. The mean values differed significantly between the two groups, according to Welch's test (t (207) = 2.42, $p < 0.05$). Thus, Cluster 1 users visited locations with tags more often than Cluster 2 users. Furthermore, the visit frequency of Cluster 2 users varied more widely than that of Cluster 1 users, according to the sample standard deviations.

Table 5. Fundamental statistics for visit frequency.

Cluster	Users	Mean	Median	SSD
1	151	8453.35	7,237	5,620.22
2	58	5853.40	3,840	9,475.01

Because staying time may be associated with the visit frequency, the staying time for each cluster was also calculated (Table 6). The results indicate that the mean staying time between the two clusters are not significantly different (t (207) = −1.63, $p > 0.05$).

Table 6. Fundamental statistics for staying time (second).

Cluster	Users	Mean	Median	SSD
1	151	3,085.05	2,375	2,357.48
2	58	3,715.70	3,353	2,785.39

(B) Mean and percentage of visit frequency by zone

Table 7 shows the zone-wise mean visit frequencies. Zones that exhibited significantly differing mean visit-frequency values between clusters were zone A (t (207) = −3.66, $p < 0.01$), zone B (t (207) = 3.40, $p < 0.01$), and zone C (t (207) = 5.12, $p < 0.01$). These results clearly show that Cluster 1 users visited the general zones (B and C) more often than Cluster 2 users, whereas Cluster 2 users visited the research zone (A) more often than Cluster 1 users. However, the standard deviations are high for both clusters; this indicates that visiting behaviors differ substantially from user to user.

Table 7. Mean value of visit frequency by zone (seconds).

Cluster	Zone	A	B	C	D	E	F	G	H	I	J
1	Mean	371	5,655	1,375	643	27	210	32	64	25	52
	SSD	1,021.67	4,570.51	2,986.94	1,463.34	165.72	371.54	126.72	113.08	52.42	163.41
2	Mean	2,433	1,564	102	900	13	563	130	76	29	44
	SSD	4,203.04	8,626.50	346.16	1,619.45	53.90	2,613.78	486.36	144.10	57.03	88.28

Zone-wise visit frequencies as a percentage of the total visits by users in each cluster were also calculated (Table 8). The results show that Cluster 1 users visited zone B most frequently (66.9%), whereas Cluster 2 users visited zone A most frequently (41.6%). For comparison, Cluster 2 users visited zone B with a frequency of 26.7% and Cluster 1 users visited zone A with a frequency of 4.4%. The differences between two clusters were extremely large. Furthermore, hypothesis testing for the difference in the population proportions between the two clusters indicated significant differences for all zones ($p < 0.01$). Thus, it can be concluded that Cluster 1 users visited the general book zone more often, whereas Cluster 2 users visited the research zone more often. These behaviors reflect the cluster construction.

Table 8. Percentage of visiting frequency by zone (%).

Cluster	A	B	C	D	E	F	G	H	I	J	Sum
1	4.4	66.9	16.3	7.6	0.3	2.5	0.4	0.8	0.3	0.6	100.0
2	41.6	26.7	1.7	15.4	0.2	9.6	2.2	1.3	0.5	0.7	100.0

4.3.3. Analysis of the questionnaire responses

The number and percentage of users providing each response were calculated for each cluster to characterize the behavior and attributes of each group. Hypothesis testing for the difference in the population proportions for all questionnaire items between the two clusters was conducted. The questionnaire items that revealed statistically significant differences between two clusters were in the “whether they borrow materials” and “whether they sit on a seat or sofa” . These results are shown in Table 9 and Table 10.

Table 9. Cluster-wise borrowing behavior.

Cluster	Borrowers		Non-borrowers		Sum	
	Num.	%	Num.	%	Num.	%
1	101	66.9	50	33.1	151	100.0
2	18	31.0	40	69.0	58	100.0

Table 10. Cluster-wise sitting behavior.

Cluster	Sitters		Non-sitters		Sum	
	Num.	%	Num.	%	Num.	%
1	82	54.3	69	45.7	151	100.0
2	45	77.6	13	22.4	58	100.0

In Cluster 1, more users borrow and fewer sit compared with Cluster 2. Testing for the difference in population proportions revealed significant differences between the two clusters for each option in both questionnaire items at a 1% level of significance.

5. CONCLUSION

Clustering based on the edit distance between users' character strings, which describe their visiting paths, enabled identification of groups of users with different behaviors in the library. The results show that the users in Cluster 1 were likely to look for materials to borrow without sitting and therefore visited more points with RFID tags. These users visited the general book zones, in which most materials available for lending are located. It is suggested that these users were more similar to one another as opposed to the users in Cluster 2, as most of the users (101 out of 119) in Cluster 1 borrowed materials.

Conversely, the users in Cluster 2 were likely to look for materials and then sit down to read them. Therefore, they visited fewer points with RFID tags than the users in Cluster 1. However, these users visited the research zone much more often than the users in Cluster 1. Cluster 2 users probably visited the library to read materials or to do research and may not have intended to borrow materials. Cluster 2 users exhibited a variety of behaviors because approximately half of them did not borrow any material. By combining these results with the finding that almost 70% of the users visited the library once a month, it can be estimated that these behaviors are characteristic of habitual library users. It is expected that location identification techniques and environments for collecting location information in libraries will improve in the future and that more studies will analyze location information data; this will lead to new insights about user behavior in libraries.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP26330367.

REFERENCES

Hui, Sam K., Fader, Peter S., Bradlow, Eric T. The traveling salesman goes shopping: the systematic deviations of grocery paths from TSP optimality. Marketing Science, 2009, vol. 28, no. 3, p. 566-572.

Julien, Heidi, Pecoskie, Jen, Reed, Kathleen. Trends in information behavior research, 1999-2008: A content analysis. Library & Information Science Research. 2011, vol. 33, no. 1, p. 19-24.

Julien, H., Duggan, L. J. A longitudinal analysis of the information needs and uses literature. Library and Information Science Research. 2000, vol. 22, no. 3, p. 291-309.

Larson, J. S., Bradlow, E. T., Fader, & Peter, S. An exploratory look at supermarket shopping paths. International Journal of Research in Marketing, 2005, vol. 22, no. 4, p. 395-414.

Sorensen, H. The science of shopping. Marketing Research, 2003, vol. 15, no. 3, p. 30-35.

Sugie, Noriko."Application of Radio Frequency Identification Technology for the Study of Information-Seeking Behavior in Public Libraries: A Preliminary Analysis".ASIS&T (American Society for Information Science and Technology) 75th Annual Meeting. Baltimore, Maryland (U.S.A.). 2012-10-26/31. (poster)

Sugie, Noriko. Application of Radio Frequency Identification Technology to Study on Information-Seeking Behavior of Library Users. Library & Information Science Research. 2013, vol. 35, no. 1, p.69-77.