Clustering of library users by similarity of visiting paths using location information

Noriko Sugie Email: sugie@surugadai.ac.jp
Faculty of Media and Information Resources, Surugadai University
698 Azu, Hanno-shi, Saitama-ken 357-8555, Japan

INTRODUCTION

Many studies have investigated information-seeking behavior in the context of user studies (Julien, 2000 & 2011). However, only a few studies have attempted to understand physical behavior in the library. 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 et.al, 2005). These studies showed that the RFID system is able to collect large volumes of accurate data about customer behavior.

Therefore, in 2012, I conducted a study to collect users' behavioral data at a public library. The present study performs a statistical analysis of the data acquired in 2012 to discover the information-seeking patterns of users.

METHOD

This study aims to classify library users' behavior by analyzing location information acquired using RFID-based observation methods.

Data collection

This study was conducted on the 9th floor of the Chiyoda Public Library in 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, 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 a tag (Figure 1).

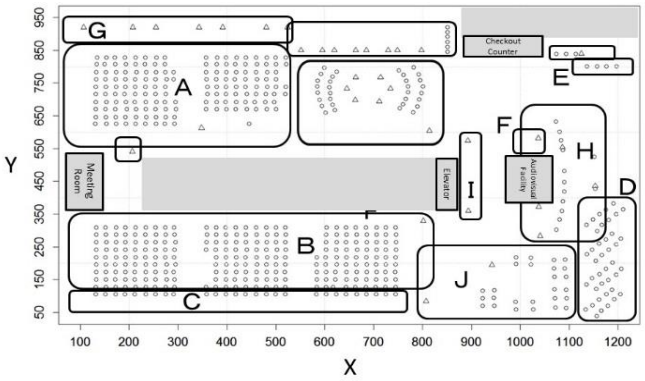
The RFID system chronologically provides the time and ID number of the RFID tags at the points visited by the users. By referencing each ID number to the bibliographic data of the library, the shelf number, position coordinates on the floor map, and zone in the library were derived.

A questionnaire survey on library usage was conducted after the users finished using the library. The questionnaire items covered user attributes, visit frequency, whether they borrow materials, and whether they sit on a chair or a sofa.

Data analysis

Analysis process

Of the data obtained, the position coordinates of the tags were used to identify groups by clustering the users' visiting paths by similarity. The position coordinates of each user were converted into alphabetic characters referring to zones. 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 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 form tags are regarded as the visiting frequencies for each point where RFID tags are located in the library.



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

Figure 1. Locations of the RFID tags and zones.

The edit distances were calculated from the character strings to express the degree of similarity among the visiting paths of users. Clustering the users' 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.

RESULTS AND DISCUSSION

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 of the users' visiting paths, and a dendrogram was obtained. 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.

Analysis of users' visiting path data
(A) Frequency of the visited points

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. The mean value for Cluster1 was 8,453 with a median value of 7,237. The mean value for Cluster 2 was 5,853 with a median value of 3,840. 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.

(B) Mean and percentage of visit frequency by zone

Table 1 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.

Zone-wise visit frequencies as a percentage of the total visits by users in each cluster were also calculated. 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%. 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.

Analysis of the questionnaire responses

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 chair or sofa" .

In Cluster 1, more users borrow and fewer sit compared with Cluster 2 (Table2). 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.

CONCLUSION

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.

It is expected that location identification techniques in libraries will improve in the future and that more studies will analyze location information data.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP26330367.

REFERENCES

Julien, H., et.al (2000). A longitudinal analysis of the information needs and uses literature. Library & Information Science Research, 22(3), 291–309.

Julien, Heidi, et.al (2011). Trends in information behavior research, 1999–2008: A content analysis. Library & Information Science Research, 33(1), 19–24.

Larson, J. S., et.al. (2005). An exploratory look at supermarket shopping paths. International Journal of Research in Marketing, 22, 395-414.

Sorensen, H. (2003). The science of shopping. Marketing Research, 15, 30-35.

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	1,021.7	4,570.5	2,986.9	1,463.3	165.7	371.5	126.7	113.1	52.4	163.4
2	Mean	2,433	1,564	102	900	13	563	130	76	29	44
	SSD	4,203	8,626.5	346.2	1,619.5	53.9	2,613.8	486.4	144.1	57.0	88.3

Table 1. Mean value of visit frequency by zone.

Cluster	Borrowing behavior						Sitting behavior					
	Borrowers		Non-borrowers		Sum		Sitters		Non-sitters		Sum	
	Num.	%	Num.	%	Num.	%	Num.	%	Num.	%	Num.	%
1	101	66.9	50	33.1	151	100.0	82	54.3	69	45.7	151	100.0
2	18	31.0	40	69.0	58	100.0	45	77.6	13	22.4	58	100.0

Table 2. Cluster-wise borrowing and sitting behavior.