GEOGRAPHICAL REPORTS
OF TOKYO METROPOLITAN UNIVERSITY 50 (2015) 73-80

SEQUENTIAL PATTERN MINING FOR ACTIVITY DIARIES: A CASE STUDY OF HOUSEWIVES LIVING IN CENTRAL TOKYO

Naoto YABE*

Abstract Sequence alignment methods have been applied in classifying activity sequences for time geographic researches. However, many sequence alignment methods cannot easily find the characteristics or “motifs” of activity patterns. A sequential pattern mining method, called Prefix Span is applied to activity sequences in this study. Prefix Span successfully determined the motifs in the activity sequences of housewives living in central Tokyo. To examine the influence of the Internet on activity sequences, the frequencies of motifs were compared between two groups: a group that used the Internet for housework, and a group that did not. The results indicated that Internet usage changed the order of activities.

Key words: Prefix Span, activity diary, Internet, time geography, Tokyo

1. Introduction

Activity diary surveys have been a common method used by researchers studying time geography. Sequence alignment methods that were originally developed to analyze DNA sequences have been applied to analyze activity diaries (Joh *et al*. 2001a, b; Kwan *et al*. 2014; Shoval and Isaacson 2007; Wilson 1998, 2006, 2008). Using sequence alignment methods, our daily activities can be represented with a sequence of characters that express each activity. Such methods can compute the similarities in the sequences and find clusters of activity patterns.

A subsequence or “motif” that characterizes activity sequences is a powerful tool for understanding activity patterns. However, sequence alignment methods sometimes experience difficulty in finding the motif of the sequences. To find the motif, sequential pattern mining methods are effective. Sequential pattern mining methods extract frequently appearing subsequences from activity sequences. Whereas several methods have proposed, the Prefix Span is considered the most efficient method in terms of its computational time and the quality of the extracted subsequences (Pei *et al*. 2001).

One of the hot topics in time geographic surveys is the Internet (Couclelis 2009). Smart phones, along with a variety of services provided through the Internet, are considered to have an

* Department of Social Studies Education, Joetsu University of Education

impact on our daily activities. Some shops sell even fresh vegetables and meats thorough the Internet and offer same-day deliveries. Such Internet services may change our activity patterns. Some studies have analyzed the impact of the Internet, with a special focus on gender differences (Ren and Kwan 2009; Ren *et al*. 2013; Schwanen *et al*. 2014). In those studies, the impact of the Internet is measured mainly in terms of the duration and timing of activities. Accordingly, there is a lack of research directed at the sequential aspect of such activities.

The present work is intended to clarify the impact of the Internet on daily activities from a sequential point of view. The motifs from Internet related activities are extracted by applying a sequential pattern mining method to activity diaries. By comparing the motifs from Internet related activities with the motifs from activities that do not use the Internet, we can discover clues that will help us to understand the impact of the Internet.

2. Data

The activity diaries were collected from housewives living in central Tokyo. Like other big cities in the developed countries, Tokyo has experienced a population growth and gentrification at the city center, from the late 1990s onward. The residents living in the city center show activity patterns that are different from those living in the suburbs, owing to better accessibility in terms of job opportunities. Because an abundance of job opportunities eases the constraints on residents, the differences in the activities of housewives are considerable relative to their suburban counterparts. To clarify the difference, Yabe (2014) conducted activity diary surveys in Tokyo's city center, and the data that Yabe (2014) collected is used in this research.

Activity diary survey

First, survey respondents were recruited through a research firm. Respondents had to meet the following four criteria to be eligible for the survey: an eligible candidate must be (1) a married woman, (2) living in central Tokyo or sub-central Tokyo, who is (3) not a student, and (4) has a child under 18 years of age. For the purposes of criterion (2), "central Tokyo" was defined as the Chiyoda Ward, Chuo Ward and Minato Ward, and "sub-central Tokyo" was defined as the Shinjuku Ward and Shibuya Ward. A total of 646 participants were identified who matched the four criteria.

Questionnaires were sent to the respondents over the Internet. Participants filled out the form online. The questionnaires asked respondents to record their activities in 30 minute intervals for the most recent weekday. The options for activities were presented in advance, and respondents could select an activity from one of 11 categories. The categories were selected by referencing previous work concerning a time budget analysis (Yano 1995). The 11 categories were as follows: (1) Sleep; (2) Meals; (3) Personal care; (4) Travel; (5) Work; (6) Housework; (7) Socializing; (8) Education/Leisure; (9) Mass media; (10) Rest; and (11) Other. For the categories of Work, Housework, Socializing, Education/Leisure, and Mass media, housewives recorded whether the activities included using the Internet. In addition to the activity diary, the employment status of working mothers and the housekeeping tasks assigned to each family member were recorded. The questionnaires were sent to participants on Tuesday, January 17, 2012 and responses were accepted until Monday, January 23, 2012. In total, 336 replies were collected; of these, 305 were considered valid replies, for a response rate of 47%.

Because the survey was conducted over the Internet, respondents were limited to housewives with Internet access. It is widely recognized that Internet surveys frequently exclude older respondents, and that this may constitute a sampling bias. In the present study, however, subjects were housewives with a child under the age of 18. As mothers with young children are considered middle aged rather than elderly, the sampling bias associated with the Internet was not thought to play a critical role.

Attributes of respondents

The majority of respondents were in their thirties (*N*=112; 37%) or forties (*N*=142; 47%). There were no respondents in their sixties. This was mainly because of the age requirement of their children. Of the respondents, 43% had children under 5 years old, and 30% had children between the ages 6 and 12. This means that almost 73% of respondents had children in preschool or primary school. Hence, it can be said that the sampling bias of Internet surveys had little effect on this study.

Among the 305 respondents, 39% were working mothers (*N*=120), and 61% were non-working mothers (*N*=185). The employment status was calculated for working mothers. The percentage of working mothers who held full-time jobs was 38%, which is considerably higher than that found among suburban residents from an earlier study (Sugiura and Miyazawa 2001). As for the place of residence, 45% of respondents lived in central Tokyo, while 55% lived in sub-central Tokyo.

3. Methods

Several sequential pattern mining methods have been proposed. Prefix Span is one such method. Prefix Span was employed in this paper because it can efficiently extract frequently appearing subsequences or motifs from activity sequences (Pei *et al*. 2001).

Prefix Span

Prefix Span extracts frequently appearing subsequences with the following procedures. First, the daily activities of survey samples are denoted as a sequence of characters. For example, if a sample reports sleeping for one hour and then eating for 30 minutes, the activity sequence of characters is AAB, where A denotes sleep and B denotes meals. The 16 categories consist of 11 categories of activities that do not involve Internet use and five categories that do. They are coded with the characters A through P: (A) Sleep; (B) Meals; (C) Personal care; (D) Work with Internet; (E) Work without Internet; (F) Housework with Internet; (G) Housework without Internet; (H) Travel; (I) Socializing with Internet; (J) Socializing without Internet; (K) Education/Leisure with Internet; (L) Education/Leisure without Internet; (M) Mass media with Internet; (N) Mass media without Internet; (O) Rest; (P) Other. Then, the activity sequences from the 305 samples are constructed. Each sequence has 48 characters, because activities were recorded in 30 minute intervals over the course of an entire day.

Second, to compute Prefix Span, the frequency of characters, or "support," must be decided in advance. For example, to extract subsequences appearing in half of the samples, the support should be set to 50%.

Third, frequently appearing subsequences are extracted, as illustrated in Fig. 1. In this example, the motifs were extracted from four samples of activity sequences, with support set at 50%. The character A, which denotes sleep, appeared at the first step in all four sequences, and it is consequently extracted as a frequently appearing subsequence because it occurs 100% of samples. The character B, which denotes meals, and F, which denotes housework, were similarly extracted, because those characters appeared in three samples. The character D appeared in only one sample, and thus it is not considered to appear frequently as a subsequence comparing with support set at 50%. As a result, we have three characters or motifs.

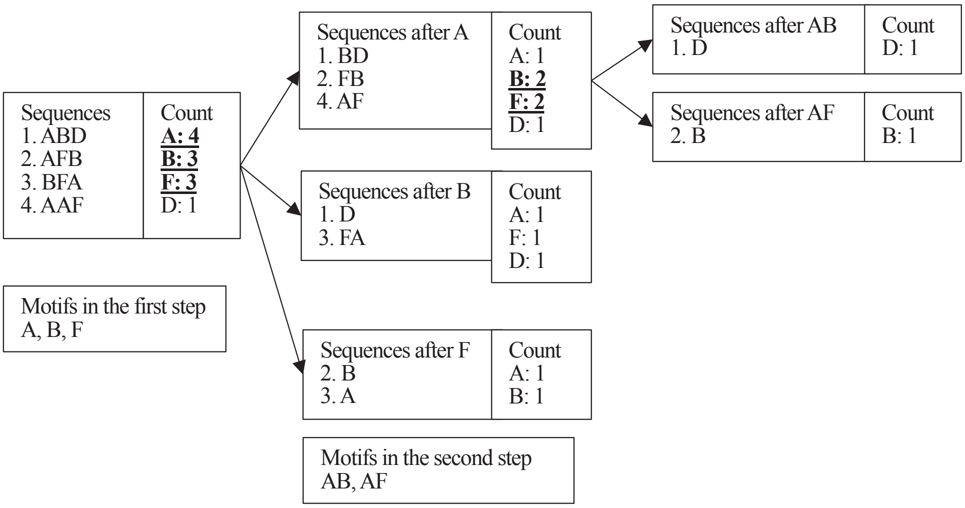


Fig. 1 Prefix Span procedures. Bold and Underline denote that the character satisfy the support of 50%.

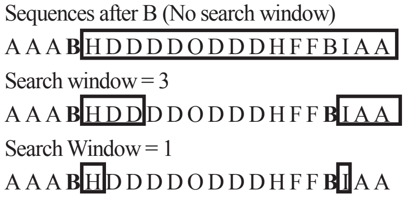
Next, we proceed by addressing the characters that follow the three extracted characters. The characters following A in the four samples are BD, FB, and AF. Any characters that appear frequently are extracted from these three sequences. In this case, B and F are extracted because they appear in two samples, satisfying 50% support. Note that when the first character in a three-character sequence is A, the motif will be two characters in length; AB and AF. The characters that follow B are considered next. The characters that follow B are D and FA. However, there are no characters that satisfy 50% support, and therefore there are no frequently appearing subsequences following B. The characters that follow extracted characters are examined in turn during the second step. Hence, AB and AF are extracted as motifs during the second step.

The third step involves analyzing the characters that follow AB and AF – namely, D and B. However, because neither character is found in more than a single sample, there are no characters that satisfy 50% support. As a result, no motifs of three characters in length are extracted. In this case we ultimately derive A, B, F, AB, and AF as motifs.

Search window

As explained in the previous section, Prefix Span extracts frequently appearing single characters in the first step before searching for the characters that follow them. Those single characters function as a prefix. Therefore, this procedure is called Prefix Span. The original procedure in Prefix Span searched for characters that followed extracted motifs without any restrictions. In other words, the characters that follow extracted motifs were analyzed, insofar as they are the remaining characters in the sequence. However, this procedure makes it difficult to interpret motifs. For instance, the motif AB can be extracted from either AB or A****B (where * is a wild card). In such cases, we cannot distinguish the proximity of B with respect to A.

A search window was introduced to correct this problem (Tono *et al*. 2004). If the search window is set to three characters, frequently appearing characters are searched within a range of three characters after the motifs extracted in the first step. This revision to the procedure makes it easier to interpret motifs. When the motif BD is extracted, for instance, the motif will be interpreted as existing when D occurs within three characters of B (Fig. 2). Likewise, if the search window is set to one, then the motif is extracted only if the characters are adjoining – that is, when they occur in sequence and unmediated (see Fig. 2).


Sequences after B (No search window)

A A A **B** H D D D D O D D D H F F B I A A

Search window = 3

A A A **B** H D D D D O D D D H F F **B** I A A

Search Window = 1

A A A **B** H D D D D O D D D H F F **B** I A A

Fig. 2 Search windows for the sequence following motif B.

4. Results

The 305 samples were divided into working mothers (*N*=120) and non-working mothers (*N*=185). Prefix Span was applied to each group, with support set at 50%. The search window was set to one in both groups.

Working mothers

For the activity sequences of working mothers, 53 motifs were extracted (Table 1). In those motifs, combinations of housework, meals, and work were the most common, with the excepting of the motif for sleep. To recognize a sequential order, there were two patterns: housework/meals after work, and work after housework/meals.

To understand the impacts of the Internet, working mothers were subdivided into two groups: the group of housewives who use the Internet for housework (*N*=32); and the group that does not (*N*=88). Then, the frequency of each motif was compared between the two groups (Nakahara and Yada 2012). As a result, the motifs that were selected show a statistically significant difference between two groups. The motif of working mothers using the Internet for housework is CGEEE. This subsequence means that 30 minutes were required for personal care, followed by 30 minutes of housework without the Internet and 90 minutes of work without using the Internet. Although Internet use was not sufficiently recurrent to constitute an extracted motif, the former group does indeed use the internet whilst performing housework. On the other hand, the motifs from working mothers who do not use the Internet for housework were G and EEGGGGB. The second subsequence means that one hour of work without the use of the Internet was followed by two hours of housework without the Internet and 30 minutes for meals. The difference between the two groups is clear: the group who uses the Internet for housework characteristically performs physical housework before going to their place of work. The group who do not use the Internet for housework, on the other hand, leaves the physical housework until after they have returned from the workplace.

Table 1 Extracted motifs from the activity sequences of working mothers

	Motifs
All samples	A, B, C, D, E, G, H, O, AA, BB, BG, CC, CG, DD, EE, GB, GG, HE, HH, AAA, BBG, BGC, BGG, CGE, DDB, DDG, EEG, GBG, GGG, AAAA, BBGC, BGCG, CGEE, DDBG, DDGG, EEGG, GBGG, AAAAA, CGEEE, DDBGG, DDGGG, EEGGB, EEGGG, AAAAAA, CGEEEE, DDBGGH, DDGGGB, EEGGBG, EEGGGG, AAAAAAA, DDGGGBG, EEGGGGB, AAAAAAAA
Group using the Internet for housework	CGEEE
Group not using the Internet for housework	G, EEGGGGB

One might consider that this difference is mainly due to the employment status, rather than because of the Internet. If a working mother is employed at a part-time job, then she can return home earlier to perform any physical housework before dinner. To clarify the effect of the employment status, the share of housewives employed at part-time jobs was compared between the two groups. The share of part-time job among the group using the Internet for housework was 34.4% (*N*=32), while the group not using the Internet for housework was 29.5% (*N*=88). However, the results were not statistically significant. Based on these findings, the difference in the activity pattern was unrelated to their employment status.

Non-working mothers

Prefix Span was also applied to the activity sequences of non-working mothers. As shown in Table 2., 43 motifs were extracted for non-working mothers. In those motifs, housework without the Internet was the most common. In particular, many motifs indicated a pattern where housework without the use of the Internet occurred after meals.

Again, non-working mothers were subdivided into two groups to examine the impact of the Internet. One group used the Internet for housework (*N*=80), and the other did not (*N*=105). The frequency of these motifs was compared between two groups, and motifs that revealed a statistically significant difference were extracted. As a result, motifs did not appear within the group who used the Internet for housework. Using the Internet to perform housework may introduce flexibility in terms of scheduling, leading to a variety of activity patterns. Therefore, no motifs were extracted from among the group who used Internet for housework. The motifs that characterizing the group who do not use the Internet for housework are as follows: BG, GB, BBG, BGG, and BGCG. Within those motifs, four motifs suggest that physical housework is performed after meals.

Table 2 Extracted motifs from the activity sequences of non-working mothers

	Motifs
All samples	A, B, C, G, O, AA, BB, BG, CC, CG, GB, GG, OG, AAA, BBG, BGC, BGG, CCG, CGG, GBG, GGG, OGG, AAAA, BBGG, BGCG, BGGG, CCGG, CGGG, GBGG, GGGG, OGGG, AAAAA, BBGGC, BGCGG, BGGGB, BGGGG, CCGGG, AAAAAA, BGGGBG, BGGGGG, AAAAAAA, BGGGBGG, AAAAAAAA
Group using the Internet for housework	No motifs
Group not using the Internet for housework	BG, GB, BBG, BGG, BGCG

5. Conclusions

This paper applied Prefix Span, a sequential pattern mining method, to activity diaries. Unlike sequential alignment methods, Prefix Span successfully found motifs in activity patterns. A difference in terms of the frequency of motifs between two groups indicated that using the Internet for housework has an impact on the sequence of activities. The motifs of activity patterns are different among those who use the Internet. For working mothers, the Internet apparently influenced this pattern by reversing the order of housework and professional work. That is, housework was done before going to work among working mothers who use the Internet – and vice-versa among working mothers who do not. For non-working mothers, the Internet apparently influences the pattern of activities insofar as there is flexibility in scheduling tasks, leading to variety of activity sequences. To capture the motifs of this group, however, larger samples would be needed, requiring further research.

Acknowledgments

This paper was presented at the annual GRECO meeting in 2013 at Tokyo Metropolitan University. We dedicate this paper to Dr. Yoshio Sugiura upon his retirement from Tokyo Metropolitan University. This work was supported by JSPS KAKENHI Grant Number 23720407, 24242034 and Fukutake Science and Culture Foundation.

References

Couclelis, H. 2009. Rethinking time geography in the information age. *Environment and Planning A* **41**: 1556-1575.

Joh, C. H., Arentze, T. A., and Timmermans, H. J. P. 2001a. A position-sensitive sequence-alignment method illustrated for space-time activity-diary data. *Environment and Planning A* **33**: 313-338.

Joh, C. H., Arentze, T. A., and Timmermans, H. J. P. 2001b. Multidimensional sequence alignment methods for activity-travel pattern analysis: A comparison of dynamic programming and genetic algorithms. *Geographical Analysis* **33**: 247-270.

Kwan, M. P., Xial, N., and Ding, G. 2014. Assessing activity pattern similarity with multidimensional sequence alignment based on a multiobjective optimization evolutionary algorithm. *Geographical Analysis* **46**: 297-320.

Nakahara, T., and Yada, K. 2012. Extracting customer behaviors from streaming data using sequential pattern mining methods. *Journal of the Japanese Society for Artificial Intelligence* **27**: 146-153.*

Pei, J., Han, J., Mortazavi-asl, B., Pinto,H., Chen, Q., Dayal, U., and Hsu, M. C. 2001. PrefixSpan: Mining sequential patterns efficiently by Prefix-Projected pattern growth. *Proceedings of International Conference of Data Engineering 2001*: 215-224.

Ren, F., and Kwan, M. P. 2009. The impact of the Internet on human activity-travel patterns: Analysis of gender differences using multi-group structural equation models. *Journal of Transport Geography* **17**: 440-450.

Ren, F., Kwan, M. P., and Schwanen, T. 2013. Investigating the temporal dynamics of Internet activities. *Time and Society* **22**: 186-215.

Schwanen, T., Kwan, M. P., and Ren, F. 2014. The Internet and the gender division of household labour. *Geographical Journal* **180**: 52-64.

Shoval, N., and Isaacson, M. 2007. Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers* **97**: 282-297.

Sugiura, Y., and Miyazawa, H. 2001. Are housewives living at Utsukushigaoka happy?: From the time use survey of the housewives in the Minami-osawa district in Tama New Town. *Notes on Theoretical Geography (Riron Chirigaku Noto)* **12**: 1-17.*

Tono, H., Kitakami, H., Tamura, K., Mori, Y., and Kuroki, S. 2004. Mining of sequential patterns with variable wildcard regions using modified PrefixSpan method. *DBSJ Letters* **3**: 61-64.**

Wilson, C. 1998. Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A* **30**: 1017-1038.

Wilson, C. 2006. Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environment and Planning A* **38**: 187-204.

Wilson, C. 2008. Activity patterns in space and time: Calculating representative Hägerstrand trajectories. *Transportation* **35**: 485-499.

Yabe, N. 2014. Time budgets of working mothers living in central Tokyo: An analysis of the impacts of the Internet. *Journal of Geography (Chigaku Zasshi)* **123**: 269-284.**

Yano, M. 1995. *Sociology of Time Budgets*. Tokyo: University of Tokyo Press.*

(*: in Japanese, **: in Japanese with English abstract)