Vol. 48 No. 1 IPSJ Journal Jan. 2007

Regular Paper

Exploratory Session Analysis in the Mobile Clickstream

Toshihiko Yamakami[†,††]

As a sub-day-scale behavior analysis, the length of sessions in multiple mobile Internet services was examined. Using mobile clickstreams with user identifiers, two analyses were performed: a preparatory study for timeout values in session identification in 2000 and a long-term observation of session lengths and clicks per session in 2001. The first study showed that 10 minutes is a suitable timeout value for the observed mobile web. The second produced inter-service comparisons and showed the effects of different mobile-Internet-specific factors. The limitations and challenges for mobile-clickstream-based session identification are also discussed.

1. Introduction

As the mobile Internet emerges, the importance of measures of success for mobile web sites is increasing. With the easy-come and easy-go nature of mobile users, it is crucial to capture their behaviors in a timely manner. This behavior has 24-hour 365-day characteristics, so a study on the sub-day scale will reveal real personal behaviors. Several factors affect how network services interact with social systems. Examples include organizational history, informal relations, and time. Of these, time is an important aspect to be explored from the perspective of technologies and social systems. Mobile-Internet-specific factors impact time-based behavior observations in mobile clickstreams. Many mobile Internet pages are short-lived in order to utilize the precious screen space. Navigation depth is restricted due to the limited visibility of the deeply nested web structure. It is important to explore the specific characteristics of the mobile Internet in order to understand the behavior of its users in the mobile context. I analyzed long-term mobile web logs to identify the characteristics of user behavior on the general sub-day scale.

2. Purpose of Research

To identify user loyalty to mobile web sites, this research aims to identify methods of using the session length of mobile Internet users, which are applicable to a wide range of mobile web sites. The stay time of a session can be used to measure the user's satisfaction with the service. By tracking a group of users, we can use methods to identify the success of the mobile web over a long span of time.

3. Related Studies

Time can play a crucial role in the analysis of web usage. Temporal data mining has been an active area of research. I have performed an interval analysis on mobile Internet web sites [1] and time zone analysis [2]. Halvey, et al. described the significance of the time of day in mobile clickstreams [3] to indicate differences in user behavior on weekdays and weekends. Navigation analyses of clickstreams have been performed by many researchers [4]. The session identification of clickstreams was discussed by Anderson, et al.[5]. There are three standard session identification methods: the standard timeout method, the reference length method, and the maximal forward reference method. Recent research has tried to capture contextual information to identify the sessions. In the mobile Internet, the web pages are short-lived and they cannot easily make use of contextual information. Typical navigation path analysis does not suit short-lived pages due to the lack of long-term persistent navigation paths. Buerklen, et al. reported that web viewing time follows a Pareto distribution [6]. Srivastava, et al. used the 30-minute timeout value commonly found in survey papers [7]. Xie, et al. also used 30 minutes (1,800 seconds) [8]. The session length depends on the type of content. Veloso, et al. reported that they used 60 minutes for the timeout value for live streaming sessions [9]. Those results led me to expect a shorter session length for mobile clickstream cases.

† ACCESS
†† Graduate School of Engineering, Kagawa University

4. Method

From the mobile clickstream, I used the timestamps of web visits with user identifiers to identify sessions. The length of each session shows the stay time, which indicates the active interest of each user. Defining *a session* as a series of web visits, I examined possible multiple timeout values and evaluated the stability of the outcomes to identify an appropriate timeout threshold value that gives maximum stability over a long-range observation. When the timeout value approaches zero, there are many sessions that consist of only one click. I assumed that the best-fit timeout value is one that makes the average session length the most stable while the timeout value is changed. The timeout values were examined on only the subday scale considering the patterns of human life and the daily content update cycle of daily news services.

The analysis system is outlined in **Fig. 1**. The raw logs are stored in monthly bases. From the raw logs, the preprocessing program picks out user identifier-based per-user visit logs and produces files for the main processing. It is implemented in PHP [10]. The main processing program is implemented in R [11].

To evaluate the timeout value for the threshold, I observed the average session length for different timeout values. The method of identifying sessions is depicted in **Fig. 2**. Each Δt_n denotes the interval between time stamped clicks on a web clickstream. When Δt_n exceeds a timeout threshold value (Δt_{th}), the sequences of commands before and after Δt_n are considered to be different sessions. In the example, only Δt_3 exceeds t_{th}, so the command sequences before Δt_3 are *session n* and the ones after it are *session n+1*.

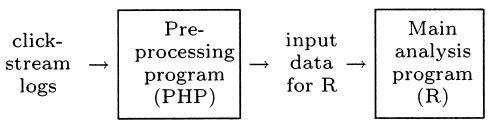


Fig. 1 Outline of the analysis system.

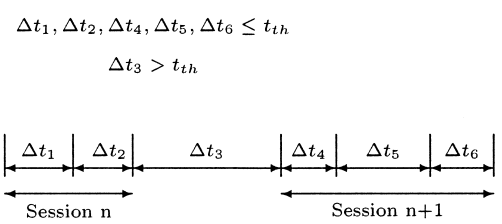


Fig. 2 Session identification using timeout values with threshold t_{th}.

I performed two exploratory analyses on mobile clickstreams. First, I performed a session length analysis, taking *timeout-value*-based session identification as explained above. To identify the timeout threshold value that suits the services, I performed a one-month analysis of data for December 2000. After this preparatory analysis, I used the timeout value to analyze the transitions of sessions over a span of time to follow the trend of session length.

5. Case Study

5.1 Data Set

The observation target was a commercial news service on the mobile Internet. The service is available from three different mobile carriers, with slightly different content menus. Each mobile carrier has different underlying network characteristics and different charging policies. The service is charged via a monthly subscription fee. The log stores the unique user identifier (UID), time stamp, command name, and content shorthand name. The services were launched from 2000 to 2001 and continue to operate now. The target service provides 40 to 50 news articles per week on weekdays. The commercial mobile service charges a monthly subscription fee to users of approximately US$ 2.5 per month. The UIDs are usually a unique string of 16 or more alphanumeric characters, e.g., "310SzyZjaaerYlb2". The service uses Compact HTML [12], HDML [13], and MML, which is a dialect of HTML. During the observation period, the Compact HTML service used packet-based networks, the MML service used connection-oriented networks, and the HDML service used both of them. The service is a business-oriented one and 90% of users were male on the Compact HTML site, where a user profile was available for some users with premium service registration. The main user segment was in the age range of 25–40. The other two services did not gather user profiles, but the user profiles were assumed to be similar considering the paid business-oriented news service characteristics.

5.2 Timeout Value Analysis

For different timeout values, I compared the clicks inside and outside the sessions. The analysis was based on the following assumptions:

- Shorter timeout values create more isolated single-click sessions,

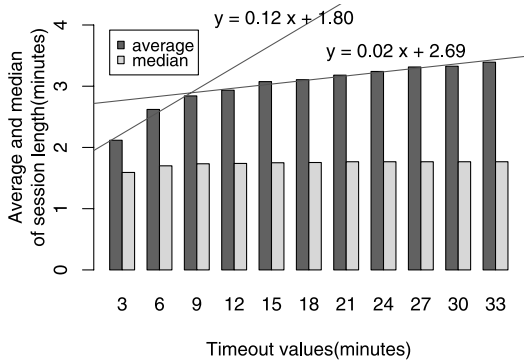
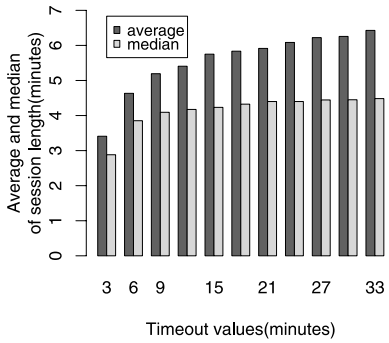


Fig. 3 Session length versus timeout values ranging from 3 to 33 minutes.

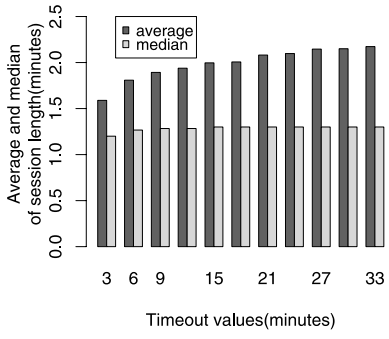
- Longer timeout values create merged sessions with session times that are longer than in reality, and
- Timeout values are in the sub-day range, e.g., less than 1,440 minutes (= 24 hours).

For the different timeout values, I compared the average and median session lengths. I examined timeout values from 3 to 1,024 minutes. Above 96 minutes, the average value grew noticeably. The median value did not change much over a wide range of timeout values. This shows that the average time increases with the timeout value. Based on this result, shorter ranges of timeout values were examined in detail. A more detailed test was performed on timeout values ranging from 3 to 33 minutes. The results are shown in **Fig. 3**. The median value was stable between 6 to 33 minutes. The average values showed two trends: a line with a steep gradient during the range of 3 to 9 minutes and one with a slight gradient from 9 to 33 minutes. The linear system approximations are drawn in Fig. 3. The lines cross at a timeout value of 10 minutes. This is considerably shorter than the values used for Internet pages aimed at desktop personal computers (hereinafter called the PC Internet) in the literature, such as 30 minutes.

The observed mobile web site is open to subscribers and non-subscribers and these two groups might be expected to have different behaviors. Therefore, to identify the differences, if any, I performed the analysis on each group. The type of subscription of a visitor is difficult to judge from the clickstreams. Clickstreams show separate subscription and non-subscription logs, so I made two groups of user identifiers, one for identifiers with any subscription log and the other for ones without a log.



(a) mainly registered user group

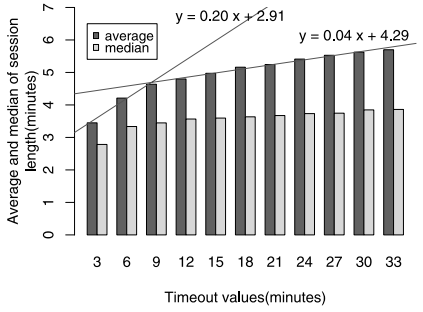


(b) non-registered user group

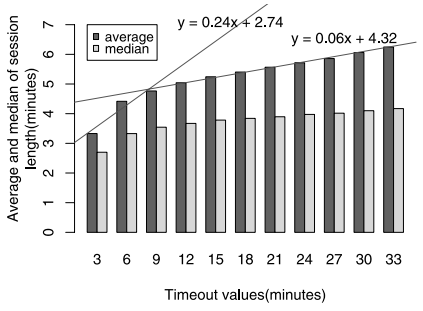
Fig. 4 Comparison of user groups in terms of average and median session lengths.

The former may include some users who were not subscribers at the time of the visit. The latter represents users without a subscription. I call the former group *mainly registered users* and the latter group *non-registered users* in the following analysis. Frequent subscription and cancellation was shown by only a negligible number of users. Therefore, I used this approximation in the following study.

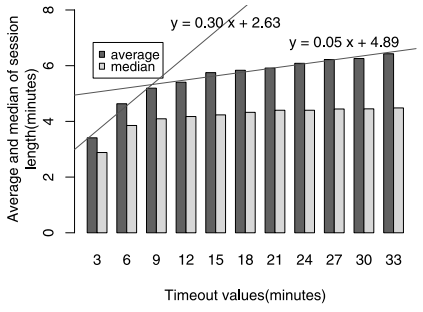
The results from the mainly registered user logs and non-registered user logs are shown in **Fig. 4**. The results in Fig. 4 (a) show that the median session length was stable in the range from 12 to 33 minutes, with a slightly increasing trend in the mainly registered user group. The differences between Figs. 4 (a) and (b) show that the two user groups exhibited different behaviors. The median and average lengths of sessions of non-registered users were smaller than the mainly registered users for all the timeout values. The median value showed little difference over a wide span of timeout value ranges. The average value showed less significant changes over the observed ranges. The non-registered users had only limited access to the content, and access to most of the paid con-



(a) an HDML site



(b) a Compact HTML site



(c) an MML site

Fig. 5 Session length analysis for mainly registered users.

tent was blocked. This led to the smaller session length in a time-scale-measured manner. It shows that the session length for clickstreams of registered users gives a clue to the session length. The following part describes my examination of the mainly registered user group focusing on the session length trend.

To verify the results for the three services, I performed a timeout value analysis on the three services in the second month after the service launch. The session analysis was performed on the mainly registered users. The results for the three services, focusing on mainly registered users, are shown in **Fig. 5**.

This confirms that all three services had similar trends. The two linear approximations from the two groups, one from 3–9 minutes and one from 12–33 minutes fit in all three cases. This supports the 10-minute timeout value as the session identification parameter for long-term observation.

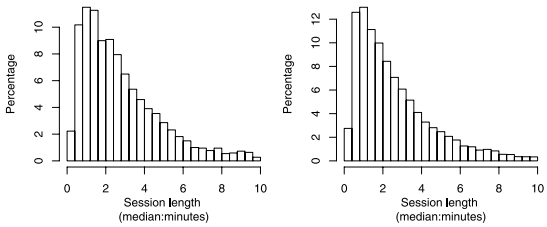
5.3 Observations from Long-term Session Analysis

Using a single timeout value for multiple services enables the inter-service comparison to identify the impacts of mobile-Internet-specific factors. I performed a long-term session-length analysis on a service for January to June 2001. To highlight the median value method by excluding exceptional data with only short-lived sessions, I used data for only clickstreams with 3 ore more sessions in a month. This did not impact any regular user behaviors. The results are shown in **Fig. 6** for a 10-minute timeout value. The distribution is a long-tail one, so the figure shows the range of 0–10 minutes to highlight the fine-grained distribution in a short timespan. Figures 6 (a) and (b) show similar distribution patterns.

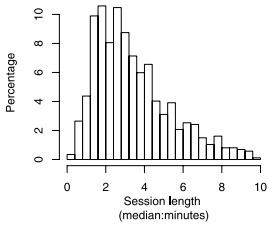
Figure 6 (c) shows a slightly different pattern with some irregularity in the distribution. Several factors affect the session behaviors:

- User profile,
- Service characteristics,
- Markup language characteristics, and
- Underlying network characteristics.

The difference in the absolute values of session lengths was assumed to be due to the number of articles and the characteristics of the connection network. The HTML and compact HTML sites had the same number of news articles, but the MML site had only two-thirds as many. Even with this difference, the MML site had a longer stay time. I consider that the connection-oriented factor explains the longer session length in the HDML and MML site. Figure 6 (c) shows a smaller number of clicks in a session, which supports the above explanation. The underlying network characteristics were different; for example, the compact HTML site was a packet-based service while the MML one was a connection-oriented service. That may partially explain the difference. The difference in markup language may also affect the service features in general. In this case, the service was provided in a one-source-multiple-presentation manner, so the HDML page contained only one card per deck, which means that interactions in HDML did not have any differences in characteristics. In this service,

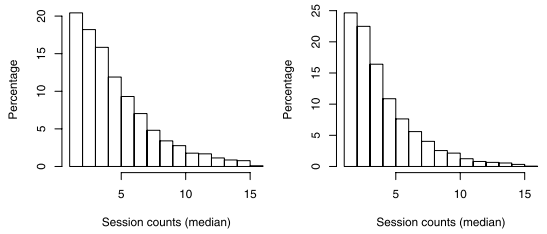


(a) an HDML site (b) a Compact HTML site

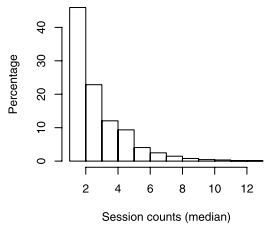


(c) an MML site

Fig. 6 Session length distribution (in the range of 0–10 minutes) for January–June 2001.



(a) an HDML site (b) a Compact HTML site



(c) an MML site

Fig. 7 Distribution of clicks per session (in the range of 0–10 minutes) in January–June 2001.

the multiple card capability in HDML was not used. The general pattern matches the Pareto distribution, but the fine-grained part showed a slight difference. This is because the mobile Internet service consists of a series of small web pages, which are different from those of the PC Internet. The irregularity in case (c) may be related to the number of users: the MML site had the fewest among the three.

The number of clicks in a session is shown in **Fig. 7** for a timeout value of 10 minutes.

The session counts showed a Pareto distribution for all three of the services, but the curves were slightly different. Cases (a) and (b) had similar curves. The numbers in the session were smaller in (c).

6. Discussion

First, the legitimacy of the obtained data needs consideration. The data set comes from 2001. The services corresponding to the data set continue to be provided today. They have witnessed the rapid evolution of the mobile Internet, Flash, PDF, SMIL, Java, and so on. However, periodical information retrieval is one of the basic user behavior characteristics to be uncovered in the behavior research. Rich media content like Java, PDF, or Flash does not easily reveal end-user behaviors in terminals. With the emergence of rich media applications, a larger number of client-side local high-level interactions are hidden from the web servers. Primitive HTML services are a good source for basic user behavior studies. In addition, it is crucial to identify the basic primitive characteristics in order to investigate the rich media contents of today and the future.

Second, the session length analysis in different carrier environments needs further consideration. The stay time depends on different factors: content structure, content length per page, content semantics, and client capabilities. In this paper, I show stable results for three different markup languages. The three services in three different wireless carriers also had different underlying network characteristics. For the same content length and for content structure from a single-source-multiple-use environment, the results were stable. This enables us to make inter-service comparisons to identify how the content length and content semantics will affect the session length.

Third, the session length analysis of mobile clickstreams made me aware of some of the challenges in this domain. Mobile clickstream session analysis involves multiple factors:

- There are multiple factors to derive shorter timeout values
  - Information is easily consumed due to the short page size,
  - Users may often leave a site due to the limited visibility: they may overlook information on the page that is not visible on their screen without scrolling or information on the next page.
- Multiple factors lead to longer timeout val-

ues
- Users carry their handsets and can easily resume their session anywhere,
- Users may carry out parallel tasks (e.g., commuting, listening to music, etc.) while maintaining a loose link to the mobile web site,
- Users may have real-world interruptions (e.g., changing trains, etc.) and resume a session after that.

The data obtained in this paper reveals the factors that shorten the timeout value. The factor in user modeling to lengthen the timeout value and maintain a loose session will be investigated in further studies. The basic result obtained here can be used to explore long-timeout factors. It also gives some clues about the linear factor for different timeout values longer than 10 minutes. The analysis of these two factors involves research studies to clarify the fundamentally multifaceted characteristics of the mobile Internet, which augments the simultaneous multiple contexts in the real world. Mobile clickstream analysis is an unexplored research field because WML1.x-based mobile Internet sites are still used in many countries. WML1.x pages consist of multiple cards in a deck, where many user clicks are absorbed in the client and not available to the servers. All the three services investigated here followed HTML-based interactions. It can facilitate the server-side session analysis and inter-service comparison in this study.

The data obtained in this paper provides a shorter span of time in session length. This gives a clue for counting session numbers that reflect the attraction that end-users feel to services provided on the mobile web. A further content semantic analysis to derive session identification could use this data to measure the improvement. Session identification gives further clues for content providers to understand why a session was terminated. The loose session with external real-world interruptions needs further study with a user model of joining/leaving real-world interruptions. A time of 10 minutes is sufficient to read one page on the mobile web. From this, I consider that the timeout value is determined by a factor that is not related to content size. The data obtained in this paper will enable content providers to tune their content page and observe what will happen to the average session length in order to improve the content. It can be used as a litmus test of whether or not end users will accept the tuned content. Mobile Internet users often leave a site and it is difficult to determine the reason for their departure. A generally applicable method like session identification can be used as the basis for measuring the dynamism of mobile Internet services.

7. Conclusion

I identified the mobile-Internet-specific session length and used a timeout comparison to identify the appropriate timeout for session identification for mobile clickstreams. I found that timeout values of around 10 minutes are appropriate for mobile clickstream session identification. Using this timeout value, I analyzed mobile clickstreams in 2000 for a 6-month period. This value is considerably smaller than the value used in PC Internet studies in the literature. There are two conflicting factors in the mobile Internet that make the session length either shorter or longer than that of the PC Internet. I described an exploratory analysis of mobile clickstream session identification and showed that the session length for mobile clickstreams is relatively short. I also showed that the median session length distribution is stable over a long span of time. The session analysis covering two different aspects of the mobile Internet exhibited a new social reality with multiple simultaneous social contexts, which is augmented by mobile Internet technology.

References

1) Yamakami, T.: Unique Identifier Tracking Analysis: A Methodology To Capture Wireless Internet User Behaviors, *ICOIN-15*, Beppu, Japan, pp.743–748, IEEE Computer Society (2001).
2) Yamakami, T.: A mobile clickstream time zone analysis: implications for real-time mobile collaboration, *Proc. KES2004* (*Volume II*), Lecture Notes in Computer Science, Vol.3214, pp.855–861, Springer Verlag (2004).
3) Halvey, M., Keane, M. and Smyth, B.: Predicting Navigation Patterns on the Mobile-Internet Using Time of the Week, *WWW2005*, pp.958–959, ACM Press (2005).
4) Ali, K. and Ketchpel, S.: Golden Path Analyzer: using divide-and-conquer to cluster Web clickstreams, *ACM KDD2003*, pp.257–276, ACM Press (2003).
5) Andersen, J., Giversen, A., Jensen, A., Larsen, R., Pedersen, T. and Skyt, J.: Analyzing clickstreams using subsessions, *Proc. 3rd ACM in-*

ternational workshop on Data warehousing and OLAP, pp.25–32, ACM Press (2000).
6) Buerklen, S., Marron, P.J., Fritsch, S. and Rothermel, K.: User Centric Walk: An Integrated Approach for Modeling the Browsing Behavior of Users on the Web, *Proc. ANSS '05*, pp.149–159, IEEE Computer Society (2005).
7) Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.-N.: Web usage mining: discovery and applications of usage patterns from Web data, *ACM SIGKDD Explorations Newsletter*, Vol.1, No.2, pp.12–23, ACM Press (2000).
8) Xie, Y. and Phoha, V.: Web user clustering from access log using belief function, *Proc. K-CAP'01*, pp.202–208, ACM Press (2001).
9) Veloso, E., Almeida, V., Meira, W., Bestavros, A. and Jin, S.: A hierarchical characterization of a live streaming media workload, *Proc. 2nd ACM SIGCOMM Workshop on Internet measurment*, pp.117–130, ACM Press (2002).
10) The PHP Group: PHP Hypertext Processor, available at http://www.php.net/ (2003).
11) R Development Core Team: *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria (2005). ISBN 3-900051-07-0.
12) Kamada, T.: Compact HTML for Small Information Appliances, W3C Note, 09-Feb-1998, Available at: http://www.w3.org/TR/1998/NOTE-compactHTML-19980209 (1998).
13) King, P. and Hyland, T.: Handheld Device Markup Language Specification, submission to W3C Note 09 May 1997, Available at: http://www.w3.org/TR/NOTE-Submission-HDML-spec.html (1997).

(Received May 31, 2006)
(Accepted November 2, 2006)
(Online version of this article can be found in the IPSJ Digital Courier, Vol.3, pp.14–20.)

Toshihiko Yamakami was born in 1959. He received his M.Sc. degree from the University of Tokyo in 1984. He is a Senior Specialist, Product Strategy Management, ACCESS. He is engaged in international standardization. Prior to joining ACCESS in 1999, he worked for NTT Laboratories in research and standardization. He was Chair of ISO SC18/WG4 Japanese National Body, IPSJ Groupware SIG vice-chair, W3C XHTML Basic Co-editor, and WAP Forum WML 2.0 Editor. He has been a Guest Professor at Tokyo University of Agriculture and Technology since 2005. He received the IPSJ Yamashita Award in 1995. He is a member of IPSJ and the Association of Computing Machinery.