IPSJ Journal Vol. 49 No. 7 2493–2499 (July 2008)

Regular Paper

A Stream-mining Oriented User Identification Algorithm Based on a Day Scale Click Regularity Assumption in Mobile Clickstreams

Toshihiko Yamakami[†1,†2]

The mobile Internet is characterized by "Easy-come and easy-go" characteristics, which causes challenges for many content providers. The 24-hour clickstream provides a rich opportunity to understand user's behaviors. It also raises the challenge of having to cope with a large amount of log data. The author proposes a stream-mining oriented algorithm for user regularity classification. In the case study section, the author shows the case studies in commercial mobile web sites and presents that the recall rate of the following month revisit prediction reaches 80–90%. The restriction of the stream mining gives a small gap to the recall rates in literature, but the proposed method has the advantage of small working memory to perform the given task of identifying the high revisit ratio users.

1. Introduction

The mobile Internet is growing and now represents a mainstream of the Internet access in Japan after only 8 years since its launch in 1999. Users of services based on IP internet access in Japan reached 86.5 million at the end of September 2007. It is 87.1% of the total of wireless subscribers. The mobile Internet is one of the leading edges of the Internet services and offers a 24-hour availability. The detailed study and modeling of temporal aspects of the mobile Internet is on a hot research agenda. The author proposes a method for identifying the revisiting users based on the assumption that only part of the clickstream data is available at a time. The author provides a method for identifying users with a monthly scale high revisit ratio from mobile clickstreams. The author evaluates how accurately the method predicts the user revisit ratio in the following month.

2. Backgrounds

2.1 Purpose of Research

The purpose of this research is to provide a method for identifying the users with a high revisit ratio on a monthly scale within the restrictions of stream mining, e.g., limited resources of computing and storage.

2.2 Related Works

Mobile clickstream analysis is a new research field. Time zone analysis was done by Yamakami[1]. Halvey presented the significance of the time of the day in mobile clickstreams[2]. Extensive studies are done in the PC Internet clickstreams for e-commerce data mining[3]. The author explored how to use the user identifier provided by wireless carriers in the mobile clickstream analysis[4]. Methodological challenges to understand the use of mobile devices and applications still exist[5]. It is also applied to the mobile clickstream by Halvey[6].

Mining data streams is a field of growing interest due to the importance of its applications and the dissemination of data stream generators. The research on continuously generated massive data caught researcher's attentions[7)–9]. Many researchers expect this trend will reach the mobile Internet very soon. In data stream mining, it is difficult to use a generic learning algorithm because it is difficult to store the entire data stream. To cope with this constraint, the domain-specific knowledge is explored. In this paper, the author uses behavior-based knowledge.

The mobile content outside Japan is still mainly dominated by SMS, a messaging service, not a web-based service. Therefore, the international literature on mobile stream mining is still at an early stage to our best knowledge.

The mobile content providers seek to identify the methods to determine users that are regular on a monthly scale, to cover the mobile-specific monthly-subscription service model. The author presented a method for evaluating monthly scale user regularity using mobile click intervals[10),11]. This clickstream uses visit intervals. Visit intervals need an extensive computing while preserving the exact order of user visits, which does not fit mobile stream-mining constraints.

The originality of this paper is to provide a method for identifying users with a high revisit ratio on a monthly scale under the constraint of stream mining.

[†1] ACCESS
[†2] Graduate School of Engineering, Kagawa University

2.3 Revisit Ratio

The author uses *revisit ratio* to measure the degree of monthly scale regularity for a specific mobile web site. This measure is used to assess the user's rating of the site. For any given month m, the revisit ratio $R(m)$ is defined as follows where $U(m, condition)$ represents users with access to any URL within the given web site in a month m, satisfying *condition*:

$$R(m, condition) = \frac{\mid U(m, condition) \cap U(m+1) \mid}{\mid U(m, condition) \mid}$$

When condition is *all*, $U(all)$ denotes all the users with access to any URL in a given web site in the month. When condition is *multiple days*, $U(multipledays)$ (or $U(md)$) denotes all users with access to any URL in a given web at least two days in that month. When the month is explicitly stated (e.g., in the tables), the month m can be omit.

3. Stream-mining Oriented Algorithm

3.1 Requirements

Data stream mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities.

In the mobile environment, a user log sequence can be scattered to multiple distributed servers. In such an environment, they need a method that can be performed in a one-path, while the strict preserving of the arrival order of clicks is not required. One-path means that the log server can deal with the logs regardless of the order in exactly one sequential scan as far as each log is identified with a correct time stamp. Many classification methods do not meet this requirement because they need a multi-path processing on the entire data set. The interval-related methods in literature are difficult to apply.

Mobile services run in a 24-hour basis. They require distributed computing to enhance the robustness especially for 24-hour operations. The 24-hour operation introduces constraints on log-aggregation and log processing. When the number of users reaches millions, it induces a significant constraint on log analysis systems.

In this environment, the one-path algorithm without any assumption of the order of arrivals is needed. The servers are distributed and log data are sent to the analysis system, but the analysis system cannot wait until all the data are collected.

Also, the storage per user should be minimal. The mobile service subscribers are given unique user identifiers from wireless carriers. It is crucial to provide a unique user identifier-based algorithm to identify user or service characteristics. Usually, it is 16 to 32 alphanumeric characters. This needs certain storage. It is possible to compress the user identifier (UID) or use hash values in order to meet the limited storage requirement. In Japan, the largest carrier has 48 millions of mobile Internet users. Each carrier utilizes a different user identifier scheme, therefore, it is intuitive that the analysis carried out is specific to each carrier. In this case, we can use 3-byte (24-bit) or 4-byte (32-bit) hash value to store each user identifier. A 24-bit value may be insufficient for tens of millions of users like in the case of Google. A 4-byte hash value is considered to be safe for any worldwide services. They can use something like 20-bit, but byte-boundary based systems are favored for their ease of operation and high performance without bit-wise manipulation. Under the assumption that each user's identification information needs 3 or 4 bytes, it is safe to set as goal the use of 1 byte per user in a stream-mining algorithm to identify the users with a high revisit ratio.

Also, the method needs to cope with the content provider's requirement of identifying very loyal users. It means the method should be applicable not only to general users but also to subscribed users. Subscribed users show more regular behavior than non-subscribed. It means the task of identifying high revisiting users among subscribed ones is more difficult compared to the general case.

3.2 A Day-count Based Algorithm

Some methods were proposed, but all methods relying on click intervals cannot be used in a mobile stream-mining oriented environment. The author performed a preliminary study to identify monthly scale high revisiting users in mobile clickstreams. It was shown that the users who revisit after a long span have a high probability to revisit the site in the following month. From this observation, the author proposes a day-count method depicted in **Fig. 1**. This algorithm needs 6-bit for the storage of the previous visit and 1-bit for marking regular

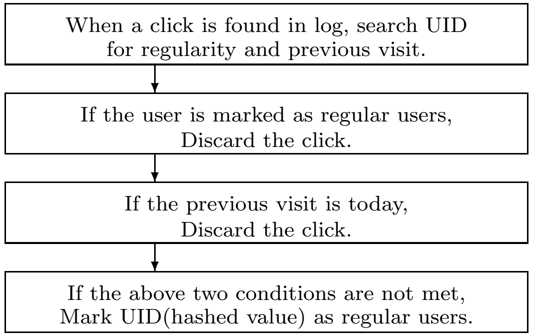
When a click is found in log, search UID for regularity and previous visit.
If the user is marked as regular users, Discard the click.
If the previous visit is today, Discard the click.
If the above two conditions are not met, Mark UID(hashed value) as regular users.

Fig. 1 A day-count algorithm to perform a unit operation on a new click log.

b1–b32	b33–b38	b39	b40
user identifier hash value	visit day	regularity mark	monthly operation mark

Note: b_n denotes bit n

Fig. 2 5-bytes memory configuration per user in bit alignment.

users. This assumes that the entire operations are performed in a monthly scale (month-independence). When the month-independence assumption is not met, another bit can be used to identify the previous month log bit to preserve the state information to the following month without any total initialization. This method identifies the users with two visits in different days in a month as regular users, with a high probability to revisit in the following month.

The memory configuration per user is illustrated in **Fig. 2**. If this "two visits" assumption is too loose, we can add further bytes to record more days to identify the regular users at the cost of the storage space, one byte per additional day, approximately.

4. Case Study

4.1 Data Set

The observation target is a commercial news service on the mobile Internet. The service is available on the three different mobile carriers, with slightly different content menus. Each mobile carrier has different underlying network characteristics and different charging policies. A monthly subscription fee is charged for the service. The log stores the unique user identifier, a time stamp, command name and content shorthand name. These services were launched from 2000 to 2001, and continue today. The target service provides 40 to 50 news articles per week on weekdays. The monthly subscription fee for users of the commercial mobile service is approximately 3 US dollars per month.

Table 1 News-access only log data set characteristics.

Carrier	Months	Clicks	Sum of Monthly U-Users	Unique Users
A	0101–0105	196,369	11,610	4,462
A	0201–0205	144,767	7,046	2,442
B	0101–0105	86,808	3,163	1,672
B	0201–0205	82,815	2,437	901
C	0101–0105	16,050	1,245	901
C	0201–0205	11,610	914	329

The UID (User ID) is usually 16 or more unique alphanumeric characters long, e.g. "310SzyZjaaerYlb2". The service uses Compact HTML [12], HDML (an early version of WML) and MML (a proprietary dialect of a subset of HTML).

In order to remove the non-news based additional services, which differ from carrier to carrier, the author filters all non-news, related transactions from logs from January to May 2001 and from January to May in 2002. The numbers of clicks and user identifiers in the log used for analysis in this paper are depicted in **Table 1**. The sum of monthly unique users (U-Users) is the sum of unique user identifiers in the five months. The number of unique users excluding the duplicates is shown in the column of unique users.

The news service has a unique property of regular information updates. Considering the combination of three (a) general, (b) service-specific, and (c) user-specific factors on regularity, this regular update pattern provides the basic scheme to identify the general mobile behavior. Other services with other update patterns can be compared to this regular update pattern in order to identify the service -specific patterns.

Table 2 Welch's t test summary.

Alternative hypothesis	True difference in means of two samples is not equal to 0
Sample 1	(0,1) vector of all users with visit days which equals or more than a threshold value here 0 means no revisit in the following month; 1 means a revisit in the following month.
Sample 2	(0,1) vector of all users with news access in the month
Tool	R's `t.test()` to test Sample 1 and Sample 2.

Table 3 Carrier-A results from January to April 2001 and from January to April 2002.

month (YYMM)	R(md) (%)	R (all) (%)	t-value	deg. of freedom	p-value significance
0101	81.98	66.59	−11.671	4088.8	0.0000 **
0102	84.47	70.93	−10.367	3845.6	0.0000 **
0103	79.73	67.11	−9.228	3952.0	0.0000 **
0104	83.30	70.93	−9.115	3650.1	0.0000 **
0201	86.22	74.65	−7.614	2631.3	0.0000 **
0202	87.23	74.72	−8.113	2481.6	0.0000 **
0203	85.86	73.15	−7.980	2463.1	0.0000 **
0204	86.57	75.19	−7.109	2323.3	0.0000 **

[Note] **: 1% confidence level *: 5% confidence level

Table 4 Carrier-B results from January to April 2001 and from January to April 2002.

month (YYMM)	R(md) (%)	R (all) (%)	t-value	deg. of freedom	p-value significance
0101	76.65	55.31	−6.173	594.5	0.0000 **
0102	81.48	56.98	−8.528	872.2	0.0000 **
0103	69.17	50.30	−6.306	923.2	0.0000 **
0104	78.78	60.52	−6.566	978.3	0.0000 **
0201	85.41	73.08	−4.511	849.7	0.0000 **
0202	89.10	75.65	−5.344	865.9	0.0000 **
0203	80.25	70.93	−3.309	907.7	0.0010 **
0204	83.24	71.90	−4.035	854.1	0.0001 **

[Note] **: 1% confidence level *: 5% confidence level

5. Results

The author evaluates how preciously the day-count method can predict the following month revisit ratio. In this case, the author extracts the users with multiple day news access logs from 2001 and 2002 clickstreams and evaluates the revisit ratio in the following month.

The author shows the statistic test for the significance of average values in two samples. Welch's t test is an adaptation of Student's t test intended for use with two samples having unequal variances. The author performed the Welch's t test for prediction and reality data. R is used to make the test with `t.test()`[13]. The test summary is depicted in **Table 2**. The (0,1) vector of all users represents the revisit reality. The test examines the proposed method's identification capability for regular users, in other words, users with multiple day visits are compared to the all users with news access in the month.

Carrier-A results from January to April 2001 and from January to April 2002 are depicted in **Table 3**. Revising ratio from the identified group using the day-count method is in the range of 79.73–87.23%. Average revisit ratio is 84.42%. Carrier-B results from January to April 2001 and from January to April 2002 are depicted in **Table 4**.

Revising ratio from the identified group using the day-count method is in the range of 69.17–89.10%. Average revisit ratio is 80.51%. The low ratio 69.17% is partly because the overall revisit ratio is 50.30% in March 2001. The 19% margin is even better than for other months. Carrier-C results from January to April 2001 and from January to April 2002 are depicted in **Table 5**. Revising ratio from the identified group using the day-count method is in the range of 78.53–87.68%. Average revising ratio is 82.61%.

All the results in Carrier-A and Carrier-B support the alternative hypothesis that the identified user group revisit ratio is not equal to all users' one with 1% confidence level. The only exceptions are February 2002 in Carrier-C case (significant only with 5% confidence level), and April 2002 (without any significance). April 2002 in Carrier-C is the only case where the alternative hypothesis is rejected.

These results express the "recall rate" for revisit ratio of identified users. The precision rate expresses the precision rate as a revisit/non-revisit user classifier. It is not covered in this paper.

Table 5 Carrier-C results from January to April 2001 and from January to April 2002.

month (YYMM)	R(md) (%)	R (all) (%)	t-value	deg. of freedom	p-value significance
0101	78.53	63.02	−3.616	414.9	0.0003 **
0102	80.79	65.89	−3.555	417.2	0.0004 **
0103	81.32	70.20	−2.698	416.9	0.0073 **
0104	83.33	70.46	−3.109	394.8	0.0020 **
0201	84.21	70.87	−3.070	353.4	0.0023 **
0202	85.62	75.66	−2.328	331.9	0.0205 *
0203	87.68	70.59	−3.917	322.9	0.0001 **
0204	79.39	69.71	−1.946	296.0	0.0526

[Note] **: 1% confidence level *: 5% confidence level

The estimation error is within the range of 13–30% mainly includes a) content unsubscription by users, b) carrier-switch by users, and c) handset-switch by users.

6. Discussion

6.1 Evaluation

The author shows the revisit ratio in the identified group is in the range of 67.17–89.10%. 18 months out of 24 months show more than 80% precision of predicting revisiting in the following. The average revisit ratio is 80.51–84.42%, equal or better than the figures in literature.

The $R(all)$ can be a revisit prediction measure. Without any classification, $R(all)$ accuracy is achievable when all users are marked as "regular". From this perspective, the proposed method achieved 10–15% (Carrier-A), 10–21% (Carrier-B), and 10–15% (Carrier-C) improvement.

The average prediction precision 80–84% is not perfect, but acceptable in many applications and comparable to the past literature using more complicated computing without any stream-mining oriented constraints. The day-count method is constructed under the constraints of limited memory and no guarantee of arrival orders. Considering the overhead the UID itself, which is 3 to 4 bytes overhead, the method is approximately optimal from the storage viewpoint. When there is a restriction that all data are byte-aligned, it is optimal.

6.2 Applications

It is important to identify what applications can use this measure to realize a value-added service in the mobile Internet. For example, the high revisit ratio in the middle range of 80% can be usable when the content providers analyze the content or user interfaces from a regular-user-centric viewpoint. The content providers can customize the user interface for the regular users to differentiate their services and maintain very loyal users. The revisit ratio can be used as a litmus test to measure the effectiveness of the new services or new user interfaces. It is difficult to capture the users' feedbacks in the mobile Internet because the user interface is limited and the user does not want additional input for service feedback. The automatic method with clickstreams is usable in this context.

The identification of users with a high revisit ratio can be utilized in many mobile applications including user interface customization and content evaluation. Content providers can customize the user interface to improve services and benefits for very loyal to the site. Content providers can also use the navigation patterns of loyal users to identify the key services and content on their site. Content providers can introduce new user interfaces or new services and evaluate them using the behavior changes among loyal users. The mobile Internet has "easy-come and easy-gone" characteristics, therefore, such applications to get an indirect feedback from users are valuable.

6.3 Limitations

This result was obtained for a news service. It may be bound to service-specific characteristics. This study is limited to one application provided in three wireless carriers. Cross-service comparison is for further studies.

The characteristics could be also user-segment-specific. For example, users' profile obtained in 2000 showed that 90% of users were male. Also, the men in their 20's and 30's represented majority of the users. This bias could impact the obtained result, for example, this pattern could be only applicable to business persons.

The data were obtained from mobile clickstreams in 2001. The comparisons with the latest mobile clickstream are needed to further verify the result.

It should be noted that the periodical update of content in a day is a basic pattern in mobile services, which are unchanged over a long span of time. Technologies like Java, AJAX, PDF, FLASH continue to evolve, however, the human behaviors like periodic query of the most recent information persist. The economic system change like flat rate could affect the behavior, however, the number of content updates of the news service per day is consistent. The data obtained in 2001 are relevant as long as this basic service pattern persists, even under a flat rate system. The services observed in 2001 are still commercially available after 7 years.

7. Conclusion

Identifying loyal users is crucial in subscription-based mobile content business. The mobile Internet is constantly challenged by the rapid growth of users. The large scale distributed log needs stream-mining oriented algorithm to evaluate user's regularity. The author proposes a day-count method, which is a multi-day assumption based algorithm to identify monthly scale revisiting users in mobile services. The method deploys a one-path execution regardless of the order of log arrivals from distributed servers. The author shows the revisit ratio recall percentage is 67–89%. 18 months out of 24 months show more than 80% recall rate of the following month revisiting prediction. As the mobile Internet penetrates into everyday life in a massive scale all over the world, it is important to coin and polish mobile stream-mining oriented algorithms. The author proposes a novel method based on the mobile user behavior observations. The 24-hour mobile user behavior will expose more human behavior patterns and will be used for creating efficient stream-mining oriented algorithms. This is the first step toward such integration engineering on the behavior and systems.

References

1) Yamakami, T.: A mobile clickstream time zone analysis: implications for real-time mobile collaboration, *Proc. KES2004*, Vol.II, Lecture Notes in Computer Science, Vol.LNCS 3214, pp.855–861, Springer Verlag (2004).

2) 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).

3) Lee, J., Podlaseck, M., Schonberg, E. and Hoch, R.: Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising, *Data Mining and Knowledge Discovery*, Vol.5, No.1-2, pp.59–84 (2005).

4) 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).

5) Hagen, P., Robertson, T., Kan, M. and Sadler, K.: Emerging research methods for understanding mobile technology use, *Proc. 19th Conf. of SIGCHI of Australia* (*OZCHI 2005*), pp.1–10 (2005).

6) Halvey, M., Keane, M. and Smyth, B.: Time based patterns in mobile-internet surfing, *CHI'06*, pp.31–34, Springer Verlag (2006).

7) Gaber, M.M., Zaslavsky, A. and Krishnaswamy, S.: Mining data streams: A review, *ACM SIGMOD Record*, Vol.34, No.2, pp.18–26 (2005).

8) Jiang, N. and Gruenwald, L.: Research issues in data stream association rule mining, *ACM SIGMOD Record*, Vol.35, No.1, pp.14–19 (2006).

9) Gaber, M.M. and Yu, P.S.: A framework for resource-aware knowledge discovery in data streams: a holistic approach with its application to clustering, *Proc. SAC'06*, pp.649–656 (2006).

10) Yamakami, T.: Regularity Analysis using time slot counting in the mobile clickstream, *Proc. DEXA2006 workshops*, pp.55–59, IEEE Computer Society (2006).

11) Yamakami, T.: An exploratory analysis on user behavior regularity in the mobile Internet, *Proc. KES2006*, Part III, Vol.LNAI 4253, pp.143–149, Springer Verlag (2006).

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) 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.

(Received October 11, 2007)
(Accepted April 8, 2008)

(Original version of this article can be found in the Journal of Information Processing Vol.16, pp.93–99.)

Toshihiko Yamakami was born in 1959. He received his M.Sc. degree from the University of Tokyo in 1984. He received his Dr. (Eng.) degree from Kagawa University in 2007. He is a Senior Specialist, CTO Office, 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.