Intraday-scale Long Interval Method of Classifying Intramonth-Scale Revisiting Mobile Users

Toshihiko Yamakami

ACCESS, 2-8-16 Sarugaku-cho, Chiyoda-ku, Tokyo, 101-0064 Japan,
e-mail: Toshihiko.Yamakami@access-company.com

Abstract Penetration of the mobile Internet has increased its visibility worldwide. This enables analysis of detailed time-dimensional user behavior data. It also increases the industry need to identify and retain mobile users with strong loyalty to a particular mobile Web site. The author proposes an intramonth-scale revisit classification method for identifying intramonth-scale, revisiting mobile users. The author performs a case study and the result shows that the proposed method shows 87 % classifier accuracy. The author discusses a trade-off between classifier accuracy and a true positive ratio.

1 Introduction

The mobile Internet has increased its visibility in Web-based services. A major Japanese social network service provider published a press release in 2007 revealing that access from handsets had exceeded that from PCs. The mobile technology provider's long-held dream, that the mobile Internet will play a crucial key role in the Internet, is coming true. The mobile Internet is characterized by its 24-hour characteristic. It is available for 24 hours a day. This means that it is crucial to identify user behaviors in a time dimension. Mobile Internet users are "easy-come, easy-go", therefore, content providers require a new methodology to identify mobile users with strong loyalty to their mobile Web sites. In this paper, the author proposes an intraday-scale long interval method of identifying intramonth-scale, revisiting users.

Please use the following format when citing this chapter:

Yamakami, T., 2008 in IFIP International Federation for Information Processing, Volume 286, Towards Sustainable Society on Ubiquitous Networks, eds. Oya, M., Uda, R., Yasunobu, C., (Boston: Springer), pp. 27–36.

2 Purpose and Related Works

2.1 Purpose

The purpose of this research is to coin a new methodology for identifying intramonth-scale revisiting mobile users.

2.2 Related Works

Mining data streams is a field of increasing interest due to the growing importance of its applications and the dissemination of data stream generators. Research dealing with continuously generated, massive amounts of data quickly caught the attention of researchers in recent years [2] [4].

Web mining has become a hot research topic as services on the Web have emerged in the last decade [10] [5]. However, many techniques used for the PC Internet cannot be applied to the mobile Internet because the lifetime of each item is short due to screen size limitations.

Mobile clickstream analysis is a relatively unexplored field of research because SMS or WML1.3-based mobile Internet sites are still used in many countries. The WML deck consists of multiple cards, where many user clicks are absorbed by the client and are not available to the server.

The dynamics and volatility of mobile Internet services prevented long-term observational studies. Considering the fast growth of the mobile Internet, it is an important research topic to be covered. Halvey has performed the first large-scale mobile Internet analysis. He has reported a positive relationship to the day of the week in the mobile clickstream [3]. Church has performed sessions and queries analysis of mobile Internet searching with large real-world data [1]. The author conducted a regularity study on the mobile clickstream and reported 80 % accuracy in users that revisited the following month using statistical data on regularity [6] It showed an 80 % true positive ratio for regularity in long-term mobile web access [8]. However, this research covered only day-scale behavior.

The author proposed an early version of the time slot method, to identify regular users with a long interval of intraday Web visits [9]. The method was coined on the conjecture that the users who come to a web service twice in one day tend to return to the service in the following month [7]. From the empirical results, it appeared to be a promising method. The method identifies a regular user with an explicit partition among the active time slots. The limitation of the previous study is that the case studies were applied only to pay-per-month services. Services requiring subscription without a fee are not covered in past literature. These services have a large degree of volatile users and have an increased need to identify the users with strong loyalty. The originality of this research is to verify the method using services that do not charge content fees.

2.3 A Regularity-oriented Service Model

The author proposes a regularity-oriented service model as illustrated in Fig. 1 [7]. The service consists of two parts: pattern identification and service customization. This paper deals with the former one. The possible customization services are illustrated in Table 1. These services can be used to enclose regular users or to improve content based on the navigation patterns of regular users.

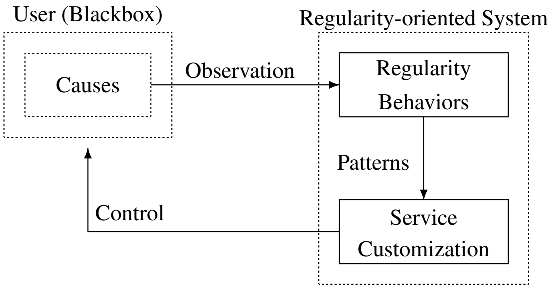

Fig. 1 Regularity-oriented service model.

Table 1 Customization services.

service	description
Menu customization	reflecting user patterns
Content rating & ranking	rating based on the navigations of regular users
Classified user service	additional service to enclose regular users

Regularity is important not only in subscription-based services, but also in mobile advertising. IDC predicts that mobile advertising will prosper in the next 5-10 years. Unique mobile-specific features in advertising include the mind-share of the Web site driving users to advertisements. Measurement of user mind-share can become an additional new feature for Internet advertising.

3 Long Interval Method

3.1 Assumptions

Past literature indicated the assumption that users who revisit a mobile Web site after a long interval within one day will have a high probability of revisiting the site in the following month.

In this paper, the author attempts to propose a method based on the assumption that users who revisit a mobile Web site after a long interval within the first visit day of the month will have a high probability of revisiting the site on other days within that month.

This assumption is based on a hybrid approach that combines quantitative measures in mobile clickstream mining and behavior modeling of mobile users.

3.2 Paid Services

The author proposes a long interval method for paid services as follows: [7]

Key Idea When a user returns to a particular web site after a long interval within a day, it is likely that he/she will return to the web site in the following month.
Long Interval Method: A method of classifying users using a long interval within one day from their click logs.
Regular Monthly Users: Group of users who will revisit the particular Web site in the following month.
Regularity classification: Two classifications, “regular”(positive) or “non-regular” (negative).
Regularity true/false evaluation: Whether a user marked as “regular” will revisit a specific mobile site in the following month.

The processing flow of classification is depicted as follows:

1. For month(m), take a sequence of clicks from clickstream for each user.
2. For each sequence, calculate the interval sequence.
3. For each interval sequence, filter out long intervals using a threshold value to identify intervals during one day.
4. Use the range-threshold for a long interval to classify it as positive (interval exists in the predefined range-threshold) or negative (no interval falls within the range-threshold).

The method provides an 88–92 % true positive ratio, however, it shows poor classifier accuracy.

3.3 Non-paid Services

Measuring user regularity is difficult for non-paid services. The author proposes a simple long interval method to measure the probability of revisiting within the same month.

First Visit Long Interval Method: A user is marked as "regular" (positive) when the user returns to a particular mobile Web side on his/her first visit day.
Regularity classification: Two classifications, "regular"(positive) or "non-regular" (negative).
Regularity true/false evaluation: Whether a user marked as "regular" will revisit a specific mobile site by the end of the month.

The processing flow is as follows:

1. For month(m), take a sequence of clicks from clickstream for each user.
2. For each sequence, calculate the interval sequence.
3. Take the first visit day interval, which is the interval before the predefined threshold value (e.g. 24 hours) in the sequence.
4. Use the range-threshold for the long interval to classify it as positive (interval exists in the predefined range-threshold) or negative (no interval falls within the range-threshold).

3.4 Classification Procedure

The classifier determines for each user whether more than the threshold value is found in the first visit day. If a long intraday interval is found, the user is marked as "will revisit", otherwise, that user is marked as "will not revisit".

The accuracy of the classifier shows the number of users in true positive and true negative classifications divided by the total number of users. The true positive ratio shows the number of users who revisited after a day-scale interval (in this case study, one day), divided by the total number of users that are marked as "will revisit". The false positive ratio shows the number of users who did not revisit after a day-scale interval divided by the total number of users that are marked as "will revisit". The true negative ratio shows the number of users who did not revisit after a day-scale interval divided by the total number of users that are marked as "will not revisit". The false negative ratio shows the number of users who revisited after a day-scale interval, divided by the total number of users that are marked as "will not revisit".

4 Case Study

4.1 Data Set

The subject of observation is a commercial car information service on the mobile Internet. This service also provides a car-auction service. The service was launched in June 2000. In June 2000, after filtering out non-mobile user identifiers, 514,180 clicks were associated with mobile user identifiers. 43,562 unique user identifiers were logged.

Basically, the service was converted from a PC Internet car information service to Compact HTML, a dialect of HTML for the mobile Internet. Although it is a commercial service, it does not charge a fee. Since it is a specialized service for a specific domain, it has a specific user segment. In relation to the nature of the accompanying auction service, the web traffic varied from month to month. The peak month and bottom month volume fluctuated at a ratio of one to nine, and had one to three volume fluctuations during the period of June 2000 to May 2001.

The service provided a used car value search, car purchase requests, a used car market value search, and a car auction service.

The service is up and operating now, however the service menus have changed over the span of time. The recent clickstream logs were not available for this research.

4.2 Results

The classification accuracy of the commercial mobile service in June 2000 is depicted in Table 2. The classifier accuracy (including both true positive and true negative), true positive, false positive, true negative, and false positive ratios are shown in the table. The internal threshold was set in the range of 2–720 minutes. Revisits are evaluated based on whether a user will return back to this specific mobile site after a certain threshold value. In this case study, the threshold value is 24 hours.

As interval threshold values increase, the classifier accuracy also increases. When interval threshold values are too small, it is difficult to get good classifier accuracy because unexpected intervals will impact the accuracy.

As interval threshold values increase, the true positive ratios improve. It should be noted that the true positive ratio does not reach 50 % even with the largest threshold value, 720 minutes (12 hours). The true positive ratio is poor in this proposed method.

As interval threshold values increase, the true negative ratios slowly decrease. The difference is not big, 91.77 % with a 2-minute threshold, and 88.88 % with a 720-minute threshold.

The classifier accuracy is driven by the true negative value, therefore, the classifier accuracy cannot exceed 88.88 % with a 720-minute threshold.

This shows a poor true positive ratio. The author attributes that the reason for the poor true positive ratio to a limitation of the first visit day analysis. First visit day analysis can give a good indication of “departing” users. However, there may be other users whose long-interval revisit patterns are shown on subsequent visit days, which are not used for classification. The first visit day analysis offers an advantage in terms of processing performance because the classifier needs to deal with the first visit day only. There is a trade-off between this performance advantage and a poor true positive ratio.

Considering the trade-off between classification accuracy and a true positive ratio, it is considered appropriate to use 12 hours as a threshold value. However, there is not a great difference between 1 hour and 12 hours because the true positive ratio is still under 50 %.

In order to improve the true positive ratio, a hybrid approach is needed to compensate the true positive ratio.

Table 2 Classification Accuracy in June 2000.

interval threshold (minutes)	classifier accuracy	true positive	false positive	true negative	false positive
2	73.76 %	22.83 %	77.17 %	91.77 %	8.23 %
3	79.67 %	26.58 %	73.42 %	91.08 %	8.92 %
5	83.06 %	30.24 %	69.76 %	90.54 %	9.46 %
10	85.07 %	33.96 %	66.04 %	90.13 %	9.87 %
20	85.75 %	35.85 %	64.15 %	89.98 %	10.02 %
40	86.05 %	36.81 %	63.19 %	89.89 %	10.11 %
60	86.24 %	37.41 %	62.59 %	89.82 %	10.18 %
120	86.51 %	37.98 %	62.02 %	89.62 %	10.38 %
240	86.82 %	38.90 %	61.10 %	89.41 %	10.59 %
480	87.18 %	40.22 %	59.78 %	89.13 %	10.87 %
720	87.42 %	41.25 %	58.75 %	88.88 %	11.12 %

5 Discussion

5.1 Usefulness of a Behavior-Based Hybrid Approach

The case study demonstrates the positive result on the assumption of a relationship between multiple visits with a long interval on the first visit day and the probability of revisiting on another day within the month.

It indicates that first day behavior can be used for user classification. It is a valuable finding for exploring mobile user behavior, which is characteristically “easy come, easy-go”. This has a certain effect on performance that avoids large-scale computation because first day analysis is easier than other methods of analysis that utilize click logs with a long span of time.

This also suggests the implication that behavior modeling can improve data-intensive mobile data mining by using a high-level behavior assumption that is derived from long-term observations. The increasing size of mobile click logs provides a challenge for mobile data mining. The promising result from this hybrid approach for behavior modeling and data analysis is a step forward for further mobile data mining, especially in terms of user regularity.

5.2 Limitations

These results were obtained within the scope of a specific service. The characteristics that were encountered could be service-specific or user group profile-specific.

The data were obtained from mobile clickstreams in 2000. Comparisons with the latest mobile clickstreams are needed in order to trace the recent results.

It should be also noted that the periodical update of content during a day is a basic, unchanging mobile service pattern. The data obtained in 2000 are applicable as long as this basic service pattern persists. The services observed in 2000 are up and in commercial operation today, however, the most recent data were unavailable for this research.

5.3 Applicability

The identification of users with a high revisit ratio can be utilized for many mobile applications including user interface customization and content evaluation. It is important to identify which applications can use this measure to realize value-added services on the mobile Internet.

The revisit ratio can be used as a litmus test to measure the effectiveness of new services or new user interfaces. It is difficult to capture user feedback on a mobile Web site because the user interface is limited and users often do not want to provide additional input for service feedbacks. 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 based on behavior changes among loyal users.

Also, it is feasible for content providers to differentiate between users with high retention and others in their services. This could improve a mobile Web site’s capability to capture users.

The mobile Internet has an “easy-come and easy-gone” characteristic, therefore, applications such as these, which gain indirect feedback from users, are very valuable.

A major Japanese wireless carrier announced in a press release recently that they made their unique user identifier system available to both official and non-official carrier sites, starting in March 2008. This will leverage the applicability of the proposed method because the user identifier is available to all content providers.

5.4 Cross-service Comparison

There are distinguishable differences between the news service studied in past literature and the car information service observed in this paper. It is interesting to note that the news service in the literature reported a high true positive ratio and poor classifier accuracy. In contrast, the case study in this paper shows high classifier accuracy and a poor true positive ratio. The differences are depicted in Table 3.

Table 3 Cross-service comparison.

	news service	car information service
intra-month log volume	rather stable	fluctuating
fee	monthly fee (partially free)	free
target	paid experience	all
true positive ratio	high	low
classifier accuracy	low	high

Stable and general services are considered to have a high true positive rate using the intraday interval method, and fluctuated and specific-preference services have a high true negative ratio and a poor true positive ratio.

Although the contrast is interesting, the fundamental cause of this differentiation cannot be identified from this study alone. Further analysis of this difference remains for further studies.

6 Conclusion

The mobile Internet is characterized by “easy-come and easy-gone” users. The mobile Internet has a unique perspective on this from the dimension of time. The author proposed an intraday long interval method for identifying users with a high probability of revisiting on an intramonth scale. The case study showed classifier accuracy of 90 % with commercial data from 2000.

This method can be applied to a wide range of mobile data services, assuming only a user identifier and time stamps in clickstreams. This method can be applied in the first month of deployment because it uses intraday behavior and intramonth prediction.

This method can be utilized to distinguish users with different loyalties to a specific mobile Web site. Therefore, it enables the tracking of user clickstreams with a higher probability of revisiting in order to identify preferred content and services, which, in turn, facilitates the timely improvement of mobile Web content.

References

1. Church, K., Smyth, B., Cotter, P., Bradley, K.: Mobile information access: A study of emerging search behavior on the mobile internet. ACM Transactions on the Web (TWEB) **1**(1), Article 4 (2007)
2. Gaber, M.M., Zaslavsky, A., Krishnaswamy, S.: Mining data streams: a review. ACM SIGMOD Record **34**(2), 18–26 (2005)
3. Halvey, M., Keane, M., Smyth, B.: Predicting navigation patterns on the mobile-internet using time of the week. In: WWW2005, pp. 958–959. ACM Press (2005)
4. Jiang, N., Gruenwald, L.: Research issues in data stream association rule mining. ACM SIGMOD Record **35**(1), 14–19 (2006)
5. Liu, J., Zhang, S., Yang, J.: Characterizing web usage regularities with information foraging agents. IEEE Transactions on Knowledge and Data Engineering **16**(5) (2004)
6. Yamakami, T.: Regularity analysis using time slot counting in the mobile clickstream. In: Proceedings of DEXA2006 workshops, pp. 55–59. IEEE Computer Society (2006)
7. Yamakami, T.: A long interval method to identify regular monthly mobile internet users. In: AINA2008 Workshops/Symposium (WAMIS 2008), pp. 1625–1630. IEEE Computer Society Press (2008)
8. Yamakami, T.: A stream-mining-oriented user identification algorithm based on a day-scale click regularity assumption. Journal of Information Processing **49**(7), xxx–xxx (2008)
9. Yamakami, T.: A time slot count in window method suitable for long-term regularity-based user classification for mobile internet. In: MUE 2008, pp. 25–29. IEEE Computer Society Press (2008)
10. Zaiane, O.R., Man Xin, J.H.: Discovering web access patterns and trends by applying olap and data mining technology on web logs. In: ADL ’98. IEEE Computer Society (1998)