

Mobile Video User Revisit Analysis based on Multi-day Visiting Patterns

Toshihiko Yamakami

CTO Office

ACCESS

2-8-16 Sarugaku-cho, Chiyoda-ku, Tokyo, Japan 101-0064

Email: Toshihiko.Yamakami@access-company.com

Abstract—The mobile Internet is characterized by “Easy-come and easy-go” characteristics, which causes challenges for many content providers. Enclosing end users and increasing mind share of each service are crucial for service adoption. The 24-hour clickstream provides a rich opportunity to understand user’s behaviors. It also raises the challenge of coping with a large amount of mobile web log data. The author examines a multi-day algorithm for user monthly-scale revisiting behavior classification for mobile video users. This was used in legacy text-oriented service in the past, however, the coverage of mobile video service users is still to be covered. 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 is approximately 80 %. The restriction of stream mining gives a small gap to the recall rates in literature, but the method has the advantage of small working memory to perform the given task of identifying the high revisit ratio users.

I. 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 92.0 million at the end of September 2009. It is 83.9 % 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 still 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.

II. BACKGROUNDS

A. 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.

B. Related Works

Mobile clickstream analysis is a new research field. Time zone analysis was done by the author [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] [8] [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 a challenge to store the entire data stream. To cope with this constraint, the domain-specific knowledge is explored. In this paper, the author uses mobile 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 lies in providing a method for identifying mobile video users with a high revisit ratio on a monthly scale under the constraint of stream mining.

C. 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*:

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

		correct result	
		positive	negative
obtained classification	positive	tp(true positive)	fp (false negative)
	negative	fn(false negative)	tn (true negative)

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 *multipliedays*, $U(multipliedays)$ (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.

The classification is defined as follows:

Precision is defined as the number of positive samples identified by a classifier divided by the total number of samples identified by that classifier, and Recall is defined as the number of positive samples identified by a classifier divided by the total number of positive samples (which should have been identified). Accuracy is defined as the number of truly identified samples divided by the total number of samples.

In other words, Precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been). Accuracy is the sum of the true positives and true negatives divided by the total samples.

F-measure that combines Precision and Recall is the harmonic mean of precision and recall. General F-measure is a weighted measure, however, we use F_1 -measure where we put the same weight to recall and precision.

$$\begin{aligned}
 Precision &= \frac{tp}{(tp + fp)} \\
 Recall &= \frac{tp}{(tp + fn)} \\
 Accuracy &= \frac{tp + tn}{(tp + tn + fn + fp)} \\
 F - measure &= 2 * \frac{precision + recall}{(precision + recall)}
 \end{aligned}$$

III. STREAM-MINING ORIENTED ALGORITHM

A. 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 for load balancing.

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 require 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 required. 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 capacity. 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.

From past literature, it is known that 2-bit + resulting mark bit(1 bit) is the minimum storage per user for identifying monthly-scale revisiting users. In this paper, the author uses this storage restriction per user to identify the high revisit possibility user group.

In the machine learning domain, computing-intensive methods like SVM (Support Vector Machine) have been proposed. The computing-intensive method requires further considerations to cope with the large-scale data of mobile web logs.

B. 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 1-bit for the storage of the previous visit and 1-bit for marking regular 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

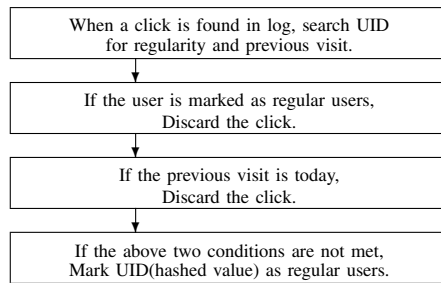


Fig. 1. A day-count algorithm to perform a unit operation on a new click log.

in different days in a month as regular users, with a high probability to revisit in the following month.

This algorithm marks a user with more than 1 day in the previous month as a revisiting user in the following month. It is straight-forward and easy to implement.

In past literature, this multi-day method provided a promising result for 2000-2002 text-based commercial news service data, with 70–90 % recall rate.

IV. CASE STUDY

A. Data set

The target of observation is a commercial video service on the mobile Internet. This service was launched in 2007.

A user ID (UID), a time-stamp, a command name and a content shorthand name are stored in the standard apache log. The content are adult movies, which a user can entertain in a stream mode or in a download mode. This commercial mobile service charges the monthly subscription fee to users, approximately 5 US dollars per month. Each channel charges this fee. There are 48 category channels. When a user subscribes a category channel with the fee, he/she watches all the content in the category channel during the month.

There are 188,926 valid unique users in March 2009 and 182,230 in April 2009. There are 16,667,364 lines of log in March 2009 and 17,708,308 in April 2009. The analysis is performed in the web log without user subscription database. From the log with URL which is accessible for only subscribed users, the approximate classification of subscription was done. From this approximation, the log contains 47,912 subscribed users, and 141,015 non-subscribed users. For subscribed users, the revisit ratio in April 2009 was 60.24 %. For non-subscribed users, it was 26.23 %. The log contains 0.5–0.7 million lines per day. The peak access hours are 22:00-02:00 around the midnight. It should be noted that the system uses Content Delivery Network for image and video delivery efficiency. The once-access image and videos are cached and there is no successive log remained in the web server side. It is impossible to access images and videos without prior web page access, therefore, this does not impact the revisit ratio analysis.

TABLE I
CLASSIFICATION PERFORMANCE FOR MARCH 2009 DATA.

	F-measure (%)	Recall (%)	Precision (%)	Accuracy (%)
Total	63.27	63.40	63.14	74.45
Subscribed	82.83	79.95	85.92	78.54
Non-subscribed	46.90	48.55	45.37	73.05

V. RESULT

The case study on March and April 2009 indicates 34.86 % users revisited in April 2009. For subscribed users, the ratio is 60.24 %. For non-subscribed users, the ratio is 26.23 %.

When the multi-day method is applied, 63.40 % of the users with more than one day visit in March 2009 revisited in April 2009. The true positive ratio is 63.40 %, and the true negative ratio is 80.32 %. The accuracy is 74.45 %.

For the subscribed users, the true positive ratio is 79.95 %, and the true negative ratio is 75.95 %. The accuracy is 78.54 %. For the non-subscribed users, the true positive ratio is 48.55 %, and the true negative ratio is 81.01 %. The accuracy is 73.05 %.

In past literature, the analysis was performed for the subscribed users. Compared to that, the true positive ratio of 79.95 % and the accuracy 78.54 % provide positive results. These results are similar to the values presented in past literature (69.17 – 89.10 %).

The summary of the classification performance is shown in Table I.

The result shows the stable accuracy from 73.0–78.9 % in all categories. The subscribed group case, it shows the promising result, 79.95 % recall rate and 85.92 % precision rate, leading to the 82.83 % for F-measure. This provides a meaningful measure to trace a monthly-scale regular users to retain the subscribed users. This result is comparable to the past literature in legacy text-based users. The result shows the monthly-scale revisiting behaviors are universal both in evolution stages (in the early mobile Internet and the recent mature mobile Internet) and in media types (text-based service and video download/streaming service). This provides an important clue to identify regular users in the mobile Internet.

VI. DISCUSSION

A. Advantage

Past literature analyzed the text-based information system in 2000-2002. The case study shown in this paper is the video-based service in 2009.

The day-count method needs only 2 bit per user to perform, which is optimally storage-efficient.

Considering this constraint, the result is promising to explain the general user monthly user behavior patterns for monthly subscription based services.

Also, this provides a benchmark result to compare other computing-intensive and resource-consuming revisit behavior classifiers.

The result shows the balance among recall rate, precision, and accuracy, especially in the subscribed group case. The

provides a positive evidence that a user returns to a mobile web site after a certain long interval has the distinguished high possibility to revisit the mobile web site in a long-term scale.

During the early stage of mobile Internet in 1999, the micro-browser-empowered mobile handset was equipped with 1 MHz CPU and 1 M byte main memory, display resolution. In 2010, it is estimated that the volume-zone mobile handsets will be equipped with 600 MHz CPU and 256 M byte main memory with flash memory 16G byte. Even with this more than 100 times expansion of device capability, the user behaviors have many commonalities, especially in regularity exposure in the mobile web logs.

Regularity is an important dimension in order to enclose end users who usually exhibits easy-come and easy-go characteristics. Regularity, especially, long-term regularity is a crucial indicator of the end user mind share of a service. In the most mobile web services, there is not much difference between top menus for royal users and ones for irregular users. This is partly because the original PC Internet did not pay much attention to regularity. There were not many business models in the PC internet depending on any of user regularity. This design principle needs to be revisited in the mobile web design. The content providers are required to pay more attention to methods for retaining regular users and for increasing regularity.

A significant percentage of mobile web traffic comes from volatile easy-come and easy-go users. Considering increased importance of communities and transaction-based business models, it is important to filter the user access according to regularity, leading to regularity-focused value-added services.

From this perspective, the proposed method is light-weight and efficient to capture 80 % of regular users in the subscribed user group.

The day count method needs only 2 bytes for processing date, which provides an advantage of intermediate storage capacity.

B. Limitation

The service specific factors or user specific factors (age, gender, et al) remain for further analysis. Also the region-specific factors (Japan) require addressing.

The fundamental user motivation analysis for attitude changes is missing in this paper. Also, the detailed content-specific user demography and usage analysis are also missing.

The network tariff for each user is ignored for simplicity in this paper.

Multiple influencing factors like service evolution stages in the region and service types are for further analysis.

The cross comparison was performed only for the legacy text news services. No recent cross-service comparison is provided in this paper.

This paper addresses a quantitative measure to trace continued use with revisits. It is crucial to coin a method for retaining user positive attitude over a long span of time. This paper does not address the concrete methods to address it.

TABLE II
MOBILE SERVICES UTILIZING REGULARITY ANALYSIS.

Service	Description
Stage-analysis	Measuring regular user ratio to identify the maturity of deployed services.
End-point analysis	Analyzing the leaving end point of non-regular users, identifying end user's discouraged points to improve retaining effect.
Content-analysis	Analyzing the content items viewed by regular users.
Menu differentiation	Providing premium top menus for regular, high loyalty users.

TABLE III
DIFFERENCES BETWEEN TRAFFIC-BASED ANALYSIS AND REGULARITY-BASED ONE.

Approach	Description	Emphasis
Traffic-based	Based on the total number of clicks.	Access volume. Accumulation effect.
Regularity-based	Based on the user-level continuity.	Per-user royalty. Long-term effect. User mind-share. Filtering out mobile-specific easy-come easy-go users.

C. Applications

The regularity analysis can be used for multiple mobile applications as illustrated in Table II.

First, an overview analysis to identify the regular user ratio transition provides an important clue to design mobile webs. It also provides a quantitative measure to improve the mobile web content in regards to the revisiting ratio.

Second, identifying regular users provides a method to identify content viewed by regular users. This also provides a clue to improve the content in regards to the revisiting ratio improvement.

Third, efficient identification of regular users provides an opportunity for differentiated top menus. A content provides can make use of this method to provide a differentiated menu for a high royalty user group.

Fourth, identifying regular users also provides identification of gone users. This provides a clue to identify leaving point and an opportunity to revise the content to prevent gone patterns.

The highlight of differences between traffic-based analysis and regularity-based analysis is provided in Table III. The traffic-based analysis relies on accumulated access volume, which has fits with volume-based business models. Examples include per-click advertisement. The regularity-based analysis relies on each-user continuity in a long time scale, which has high fits with subscription-based business models.

With a limited real estate available on a mobile handset, the mind share of a mobile service needs pursuit from content providers. Use regularity is a promising quantitative measure to capture such a mind share.

D. Implications

The proposed method provides an implication that the user revisiting behaviors have some common patterns in intra-day behaviors and in inter-day behaviors. It is fur further studies whether the user royalty has any difference between intra-day behaviors with long-term intervals and inter-day behaviors with day-scale long-term intervals.

VII. CONCLUSION

The multi-day method provides a good approximation of monthly revisit behaviors. The binary classification provides 80 % of recall rate and 85 % precision rate. The balanced deliverables of recall rate and precision rate show that the proposed method is applicable for identifying monthly-scale regular users in the mobile video services.

The proposed method requires only 1 bit for the historical memory for visits. This space-optimality is promising in the exploratory stage of user behavior analysis. Also, it is a positive result to capture the clear snapshot to accommodate logs from a large number of users.

Increased visibility of Internet-based video services leverages the mobile video service opportunities. The obtained result provides a promising clue to improve retention of mobile video service users.

Future research includes the cross-service comparisons, cross-region comparisons, and applications to make use of the proposed method.

REFERENCES

- [1] T. Yamakami, "A mobile clickstream time zone analysis: implications for real-time mobile collaboration," in *Proceedings of KES2004 (Volume II)*, ser. Lecture Notes in Computer Science, vol. LNCS 3214. Springer Verlag, September 2004, pp. 855–861.
- [2] M. Halvey, M. Keane, and B. Smyth, "Predicting navigation patterns on the mobile-internet using time of the week," in *WWW2005*. ACM Press, May 2005, pp. 958–959.
- [3] J. Lee, M. Podlaseck, E. Schonberg, and R. Hoch, "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, January-April 2005.
- [4] T. Yamakami, "Unique identifier tracking analysis: A methodology to capture wireless internet user behaviors," in *ICOIN-15*. Beppu, Japan: IEEE Computer Society, February 2001, pp. 743–748.
- [5] P. Hagen, T. Robertson, M. Kan, and K. Sadler, "Emerging research methods for understanding mobile technology use," in *Proc. of 19th conf. of SIGCHI of Australia (OZCHI 2005)*, November 2005, pp. 1–10.
- [6] M. Halvey, M. Keane, and B. Smyth, "Time based patterns in mobile-internet surfing," in *CHI'06*. Springer Verlag, April 2006, pp. 31–34.
- [7] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: a review," *ACM SIGMOD Record*, vol. 34, no. 2, pp. 18–26, June 2005.
- [8] N. Jiang and L. Gruenwald, "Research issues in data stream association rule mining," *ACM SIGMOD Record*, vol. 35, no. 1, pp. 14–19, March 2006.
- [9] M. M. Gaber and P. S. Yu, "A framework for resource-aware knowledge discovery in data streams: a holistic approach with its application to clustering," in *Proceedings of SAC'06*. ACM Press, April 2006, pp. 649–656.
- [10] T. Yamakami, "Regularity analysis using time slot counting in the mobile clickstream," in *Proceedings of DEXA2006 workshops*. IEEE Computer Society, September 2006, pp. 55–59.
- [11] —, "An exploratory analysis on user behavior regularity in the mobile internet," in *Proceedings of KES2006, Part III*, vol. LNAI 4253. Springer Verlag, October 2006, pp. 143–149.