A Time Slot Count in Window Method for Mobile Internet User Classification

Toshihiko Yamakami
CTO Office, ACCESS
Toshihiko.Yamakami@access-company.com

Abstract

The mobile Internet has become increasingly visible in everyday life. As mobile Internet penetration leverages content business opportunities, it is crucial to identify methodologies to fit mobile-specific demands, e.g. regularity. Regularity is one of the important measures to capture easy-come, easy-go mobile users, especially in subscription-based mobile services. The author proposes a method called the "time slot count in window" method (TCW-method) to efficiently capture regular users, while relaxing the explicit splits among time slots with visits. The author shows the case study result from 2001-2002 commercial service logs. Case studies conducted in 2001-2002 with commercial news service subscribers show that the relaxation in the proposed method loses 2-3 % in the true positive rate of regularity as a binary classifier. The results show high true positive rate, 88-91 % for the Carrier-A case and 86-92 % in the Carrier-C case. Experimental results show that the method is promising for identifying revisiting users under mobile-specific constraints..

1. Introduction

The mobile Internet has become a multi-faceted term covering a wide range of functions and aspects as it has deeply penetrated everyday life. More than a billion mobile handsets were shipped during 2007, the majority of them equipped with micro- (or even full) browsers. The penetration reveals a new aspect of human behavior and electronic commerce, with a large amount of access log data.

It also demands new measurements for evaluating users' behaviors in a mobile-specific context. The mobile handset has a small-sized screen that is a truly valuable asset for mobile commerce providers. From this restriction, it is crucial to increase the end-user loyalty, and to enclose them in mobile services. This leads to the importance of regularity analysis. For subscription-based mobile customers, it is important to evaluate the long-term regularity of visits rather than the total number of visits.

Challenges also come from the distributed servers that facilitate the rise of the number of mobile users. The large logs are distributed among multiple servers. It is difficult to make any multi-path web mining on this type of data.

The author proposes a new method for evaluating the regularity in mobile Internet services, and provides results to successfully distinguish long-term, regular visitors. The author uses logs of three services to evaluate the proposed method.

2. Related Works

There is a growing consensus that the mobile Internet will play an important role in the Internet in the coming future. The dynamics and volatility of mobile Internet services prevented long-term observational studies, even given this future forecast.

The author performed an interval analysis of mobile Internet Web sites [1]. The session identification of clickstreams was discussed by Anderson [2]. Halvey reported a positive relationship to the day of the week in the mobile clickstream [3]. Church performed sessions and queries analysis of mobile Internet search with large real-world data [4]. The author conducted the regularity study on the mobile clickstreams and reported 78 % to 82 % precision in users that revisited the following month using statistical data on regularity [5] [6]. It is still an active topic for researchers to study how many different types of regularity behaviors people show and how stable each behavior is over a long period of time.

Mining data streams is a field of increasing interest due to the importance of its applications and the dissemination of data stream generators. Research dealing with continuously generated massive amount of data has quickly caught the attention of researchers in recent days [7] [8] [9]. Considering the fast growth of the mobile Internet, it is an important research topic to be covered.

Mobile clickstream analysis is an unexplored research field because there are still WML1.3-based mobile Internet sites used in many countries. A WML deck consists of multiple cards, where many users' clicks are absorbed in the client and not available to the servers. In this study, all three services followed HTML-based interactions that can facilitate server-side session analysis.

The author also proposed an early version of the time slot method [10], to identify regular users with a long interval of sub-day web visits. The method was coined on the conjecture that the users that come to a web service twice in one day tend to return the service in the following month. The method was coined on the conjecture that the users that come to a web service twice in one day tend to return to the service in the following month [11]. From the empirical results, it appeared to be a promising method.

The method identifies a regular user with an explicit division among active time slots. This leads to inflexibility in setting up time slot sizes. The disadvantage of the proposed method is that an explicit division can be identified only after all access logs are analyzed. This is a considerable drawback when it is applied in a stream-mining manner. In stream mining, with the constraints of storage, it is desirable to identify the outcome in an on-the-fly manner. The past method did not match this requirement.

This paper adds inter-service comparison using log data in three different carriers to its early version.

3. Method

The author performed a preliminary study of commercial mobile Internet users using clickstream logs. The patterns obtained indicated that a user that returns to a Web site

after a certain length of time has a greater possibility of returning to the same Web site in the following month.

In order to capture this rule in an efficient method, the author proposes a method called the time slot count in window method (TCW-method). In the TCW-method, a window size is set to determine the revisiting user patterns. Usually, this window size is set to one day to capture sub-day-level user patterns. Then, the window is split into multiple time slots. The time slot size reflects service-specific characteristics. Usually, a window is split into 6, 8 or 12 time slots. All the slots have an equal length of time. The clickstream per user is distributed into these time slots. Then, the number of slots containing clicks is counted.

For example, if a user visits a web once every hour, it shows 24 visits in a day. When the window size is 24 hours, and the window is split into eight, 3-hour time slots, then each time slot contains 3 visits and time slot count is 8 (if there is more than one visit per time slot, it is ignored). The processing flow is illustrated in Fig. 1 with a time slot count threshold value n_{th}. This flow works as a binary classifier to determine whether a particular user is active in the following month or not.

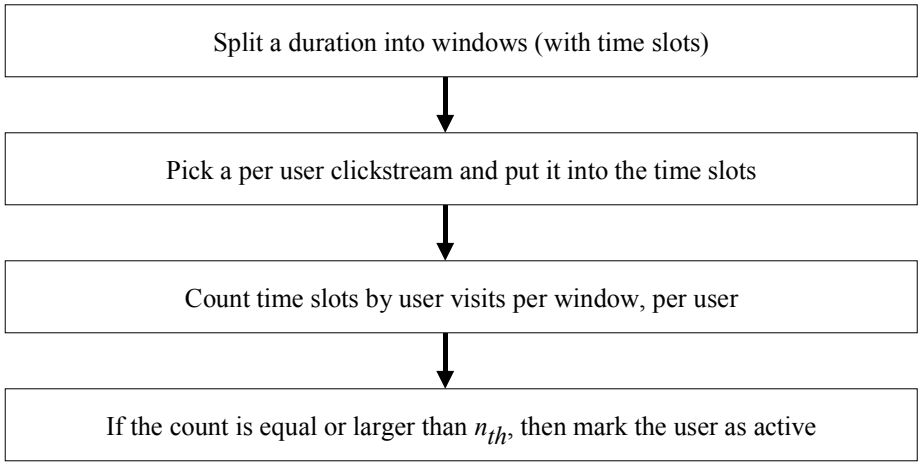

Figure 1. Processing flow of TCW-method.

An example of setting time slot bits is depicted in Fig. 2. In this example, the window size is set to 24 hours. Each window is split into 8 time slots with a 3-hour size. The clickstream for a certain user is split into windows (days). For each window, the clickstream is split into time slots. In this case, the time slot size is 3 hours, therefore, 8 time slots represents 0:00-3:00, 3:00-6;00,…, 18:00-21:00, and 21:00-24:00. 8 bits from b0 to b7 denotes a Boolean for whether there is a click during the 3-hour period. When multiple clicks fall into the same time slot, they are simply discarded. In case (a), if there is a click during 6:00-9:00 slot, then the bit b2 representing 6:00-9:00 is set to 1

because b2 was 0 before the click. In case (b), the click arrives during the same time duration, but the b2 representing the duration is already 1, so the click is simply discarded.

b0	b1	b2	b3	b4	b5	b6	b7
0	0	0	0	0	0	0	0

↓

b0	b1	b2	b3	b4	b5	b6	b7
0	0	1	0	0	0	0	0

(a) When new click is detected in the third slot b2 is set if b2 was 0

b0	b1	b2	b3	b4	b5	b6	b7
0	0	1	0	0	0	0	0

↓

b0	b1	b2	b3	b4	b5	b6	b7
0	0	1	0	0	0	0	0

(b) When new click is detected in the third slot , b2 is not changed if b2 is already 1

Figure 2. Example of setting time slot bits.

An example of counting time slot bits is illustrated in Fig. 3. When the bit is set to 1 in a window, bit 1 is counted in time slots. If time slot counts with value 1 are greater than or equal to a threshold value, then the user is marked as regular, predicting a high probability of revisiting the site in the following month. If the time slot counts fail to reach the threshold value, then the window is discarded and the next window is processed. When any window reaches the threshold value, the processing for the user is terminated, and the user is marked as regular. When the time slot counts fail to reach the threshold value in all the windows in a month, the user is marked as irregular, showing a low probability of revisiting the site the following month. In this example, the result count is 1. When this count is the largest in all windows, the user will be marked as irregular, with a low probability of visiting again the following month, when the threshold value is 2.

j

b0	b1	b2	b3	b4	b5	b6	b7
0	0	1	0	0	0	0	0

Counting bits of 8 columns.

Figure 3. Counting time slot bits.

From empirical observations, the author sets 2 as the default threshold value for the method, 3 hours for the default time slot size, and 24 hours for the default window size in this paper.

The author uses a revisit ratio to evaluate the classification of regular users. The revisit ratio R(U,m) in a month m for a group of users U is defined as follows where A(U,m) are users in U that access content (any URL in a given Web site) in the month m:

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

Where Ua is all the users that access content in the month m, R(Ua, m) represents the total revisit ratio for the month m's active users. When the active users for the month m are split into subgroups, U_1, U_2, . . . , $R(U_1,m)$, $R(U_2,m)$, . . . denotes the revisit ratio for each group of users.

This R is used for true positive ratio evaluating a binary classifier performance.

4. Case Study

4.1. Data Set

The subject of observation is a commercial news service on the mobile Internet. The service is available on three different mobile carriers, with a slightly different content menu. Each mobile carrier has different underlying network characteristics and different charging policies. The user ID (UID), the time-stamp, the command name and the content shorthand name are stored in the log.

The services were launched between 2000 and 2001, and continue to be used up to today. The target service provides 40 to 50 news articles per week, mainly on weekdays. The commercial mobile services charge a monthly subscription fee to users, approximately 3 US dollars per month. The UID 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).

The log for each carrier includes 2,390,673 lines for carrier-A, 1,591.985 lines for carrier-B, and 397,373 lines for carrier-C. The number of unique users identified in each log is 60,311 users for carrier-A, 90,291 users for carrier-B, and 13,150 users for carrier-C.

The registration records include 12,462 unique users for Carrier-A, 2,954 unique users for Carrier-B, and 1,217 unique users for Carrier-C. In order to remove the non-news based additional services, which differ from carrier to carrier, the author filters all

non-news, related transactions in logs from January to May 2001 and from January to May 2002. The data set characteristics are outlined in Table 1. Months are expressed as YYMM, for example, 0105 is May 2001.

Table 1. News-access only log data set characteristics.

Carrier	Months (YYMM)	Clicks	Sum of Monthly Unique 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	503
C	0201-0205	11,610	914	329

4.2. Result

The author performed Welch's t test for prediction and reality data. R is used to perform 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 how significant the proposed method's identification capability is for regular users, in other words, users with multiple time slot visits in a window are compared to all users that access news in the given Web site in the month.

The test examines reliability of the proposed method's for identifying regular users, in other words, users with 2 time slot of visits in a window are compared to all users that access news in the given Web site in that month. The proposed method is used for a binary classifier.

The null hypothesis is that there is no difference. Therefore, the alternate hypothesis is that there is a difference. When the null hypothesis is rejected, the alternate hypothesis is confirmed, which means that the proposed method provides a meaningful result.

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 multiple time slot visits in a window where 0 means no revisit in the next month 1 means a revisit in the next month
Sample 2	(0,1) vector of all users with news access in the month
Tool	R's `t.test()`

In the following tables, R(all) denotes the revisit ratio of all users the following month. The month is obvious from the month column, therefore, the second parameter month for R is omitted. R(TCW) denotes the revisit ratio of users identified by the TCW-method in the next month. The author performed a case study in 2001 and 2002

with 3-hour time slots in a 24-hour window. The threshold value is set to 2. The observed service is a mobile commercial news service in Japan.

The service is in commercial operation in 2007, but the recent log data were not available for this research.

The result in Carrier-A is depicted in Table 3. The t test gives a 1 % confidence level of significance in all the months under observation. The revisit ratio is 87.11 - 90.12 % range during January and April 2001, with an average of 88.74 %. It is in the 90.07 - 91.75 % range between January and April 2002, with an average of 90.00 %. It should be noted that the improvement is derived the increased revisit ratio in 2002. The revisit ratio increased from 66.59 - 70.93 %, during January and April 2001 to 73.15 - 75.19 % during January and April 2002. As time passed, the volatile users decreased and the remaining users tended to show a high revisit ratio.

The result in Carrier-B is depicted in Table 4. The t test gives a 1 % confidence level of significance in all the months under observation. The revisit ratio is 77.12 - 83.01 % range during January and April 2001, with an average of 81.24 %. It is lower than that of Carrier-A. It is because the average revisit ratio in Carrier-B(55.78 %) is lower than that of Carrier-A (68.89 %). Considering this gap, the obtained classifier performance is not inferior. It is in the 86.08 - 89.92 % range between January and April 2002, with an average of 87.19 %.

The result in Carrier-C is depicted in Table 5. The t test gives a 1 % confidence level of significance in all the months under observation. The revisit ratio is 86.30 - 92.31 % range during January and April 2001, with an average of 88.88 %. It is in the 86.08 - 91.14 % range between January and April 2002, with an average of 87.70 %.

The relaxation of explicit splits among time slots with user visits impacts the revisit ratio when a user visits for a short period of time across the boundary between two time slots. In this case, the user visits for a very short period of time, resulting in two consecutive time slot counts.

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

month (YYMM)	R(TCW)	R (all)	t-value	degree of freedom	p-value	significance
0101	87.70	66.59	-13.573	1571.6	0.0000	**
0102	90.12	70.93	-12.987	1710.7	0.0000	**
0103	87.11	67.11	-12.749	1672.0	0.0000	**
0104	90.04	70.93	-12.728	1730.7	0.0000	**
0201	91.75	74.65	-10.731	1691.7	0.0000	**
0202	91.23	74.72	-10.123	1694.6	0.0000	**
0203	90.07	73.15	-10.029	1676.0	0.0000	**
0204	90.97	75.19	-9.296	1524.8	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(TCW)	R (all)	t-value	degree of freedom	p-value	significance
0101	82.44	55.31	-6.762	258.3	0.0000	**
0102	83.01	56.98	-7.912	457.6	0.0000	**
0103	77.12	50.30	-7.985	486.8	0.0000	**
0104	82.38	60.52	-7.047	553.8	0.0000	**
0201	86.58	73.08	-4.466	573.2	0.0000	**
0202	89.92	75.65	-5.206	634.4	0.0000	**
0203	86.08	70.93	-5.225	690.1	0.0000	**
0204	86.18	71.90	-4.748	616.5	0.0000	**

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

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

month (YYMM)	R(TCW)	R (all)	t-value	degree of freedom	p-value	significance
0101	86.30	63.02	-4.634	157.8	0.0000	**
0102	90.24	65.89	-5.499	219.1	0.0000	**
0103	86.67	70.20	-3.347	162.7	0.0010	**
0104	92.31	70.46	-5.143	227.0	0.0000	**
0201	86.08	70.87	-3.014	183.7	0.0029	**
0202	91.14	75.66	-3.448	215.4	0.0007	**
0203	86.08	70.59	-3.007	190.4	0.0030	**
0204	87.50	69.71	-3.275	154.5	0.0013	**

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

In the case study, this effect is negligible, to within 2-3 % binary classifier accuracy rate, when the time slot size is a maximum of 3 hours. It should be noted that not all 2-3 % errors derive from this time slot crossing effect. The average user stay time on the mobile Internet is less then 10 minutes. Considering this factor, comparatively large time slots prevent error propagation from time slot boundary crossing patterns. The case study shows that 3 hours is sufficient to bring negligible effects. When the time slot size is smaller than 3 hours, it needs further validation tests.

5. Discussion

5.1. Advantages of the Proposed Method

The advantage of the TCW-method is that it does not depend on the final state. When the count of time slots with user visits reaches a certain threshold value, all the later clickstream for the user can be safely discarded because the later results do not impact the final identification.

The method relies on the conjecture that a user with multiple time slots of visits and a certain threshold will visit the Web site in the following month. This multiple count can be any day in the previous month. When the identification system captures multiple counts in a day, all the following clickstream can be safely discarded without impacting true positive rate accuracy. This on-the-fly nature of the TCW-method fits stream mining with its constraints on storage.

Intuitively, this method has a drawback with accuracy rate by ignoring explicit splits between time slots. The past method shows an average revisit ratio of 92.01 % from January to April 2001, and 93.88 % from January to April 2002 for the Carrier-A case. The TCW-method shows an average revisit ratio in the Carrier-A case with 88.74 % from January to April 2001, and 91.00 % from January to April 2002. It shows a loss of approximately 2-3 % true positive rate accuracy with the trade-off of on-the-fly processing capability.

There is always a trade-off between high true positive rate accuracy and recall rate. When high true positive rate is pursued, it will focus on the small group, therefore, the derived association rule can be applied to a small portion of the samples, leading to the poor recall rate. When the high recall rate is pursued, it is difficult to obtain the high true positive rate of the derived rules. Considering this trade-off, the obtained 88.74 % and 91.00 % true positive rate is acceptable for applications to exploit users' regularity. When a higher true positive rate is required for applications, it will require the sacrifice of wide coverage.

This method can be applied to a wide range of mobile applications with time stamped logs. It is a key advantage of the proposed method.

5.2. Applications

It is important to identify what applications can use this measure to realize value-added services in the mobile Internet. For example, a high revisit ratio like 90 % can be used as a measure of the impact of new services or new user interfaces. 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 acquire user feedback on the mobile Internet because the user interface is limited and the user does not want to perform additional input to give service feedback. It is desirable for content providers to differentiate users with a high potential of loyalty from others in their services. This could help improve the user retention in mobile services.

NTTDoCoMo, one of Japanese major wireless carriers, published a press release announcing that they would enable all content providers (both official and non-official carrier web sites) to use i-mode Id (their unique user identifier system). This public availability started in March 2008. In the past, the use of unique user identifiers was restricted to only carrier-approved official sites. This new carrier's movement will increase the applicability of user-identifier-based research methods.

5.3. Limitations

This result was obtained with a news service on three different carriers. The limitations include (a) it is service-specific (the result came from news services), (b) it is profile-specific (90 % users were male, most of them were in their 20's and 30's), (c) the time (2001 and 2002 data).

This bias could impact the results obtained; for example, this pattern may only be applicable to business people.

The data was obtained from mobile clickstreams in 2001 and 2002. Comparisons with the latest mobile clickstream are needed to further verify the result.

The important issues to be covered in the further studies also include:

- Trade-off between true positive ratio and recall rate,
- Impacts of different time slot size, and
- Window size different from day-scale (24 hours).

It should be noted that the periodic update of content during a day is a basic unchanged mobile service pattern that has not changed. The data obtained in 2001 is applicable as long as this basic service pattern persists.

It should be noted that the services observed are still commercially in operation today.

6. Conclusion

Mobile Internet business providers want to turn their raw data into new science, technology and revenue. One of the emerging issues in mobile Internet user clustering is based on regularity. Visit regularity reflects the mind-share of the mobile Web site, which is critical with its limited screen real estate in a small handset. The author conjectured that users that show multiple visits to a mobile Web site on any day in a given month, have a high tendency to visit the site the following month.

It is also important that the methods can be usable in a stream-mining-oriented environment. Considering the stream-mining requirement of the mobile Internet, the author relaxed the time slot count method, without the restriction of an empty slit among time slots. The proposed method can work in the not-order-preserving log arrivals in the evaluation environment, which is common in large-scale mobile services.

Empirical observations in 2001-2002 with commercial news service subscribers show that this relaxation loses 2-3 % accuracy in measuring true positive rate of regularity, but still obtains 88-91 % accuracy in the Carrier-A case. The Carrier-C case also showed similar high true positive ratio ranging from 86-92 %. In the Carrier-B, the ratio in 2001 shows the relatively low true positive ratio ranging from 77-83 %. However, the true positive rate in 2002 ranging from 86-89 % shows that the rate in 2001 is impacted from the low revisit ratio in total compared to Carrier-A and Carrier-B.

Considering the trade-off with the on-the-fly nature of coping with limited storage capacity, the author considers this method to be very promising for the mobile Internet in order to obtain regular, high-loyalty user clustering for content classification or other regularity-focused applications.

This binary classifier method is general and usable for inter-service comparisons in the mobile Internet. The method can be applied to a wide range of mobile applications that use time stamps in logs in order to capture the mobile user behaviors.

References

[1] T. Yamakami, "Unique identifier tracking analysis: A methodology to capture wireless internet user behaviors," in ICOIN-15, Beppu, Japan, February 2001, pp. 743-748, IEEE Computer Society.

[2] J. Andersen, A. Giversen, A. Jensen, R. Larsen, T. Pedersen, and J. Skyt, "Analyzing clickstreams using subsessions," in Proceedings of the third ACM international workshop on Data warehousing and OLAP, November 2000, pp. 25-32.

[3] M. Halvey, M. Keane, and B. Smyth, "Predicting navigation patterns on the mobile-internet using time of the week," in WWW2005. May 2005, pp. 958-959, ACM Press.

[4] K. Church, B. Smyth, P. Cotter, and K. Bradley, "Mobile information access: A study of emerging search behavior on the mobile internet," ACM Transactions on the Web (TWEB), vol. 1, no. 1, pp. Article 4, May 2007.

[5] T. Yamakami, "Regularity analysis using time slot counting in the mobile clickstream," in Proceedings of DEXA2006 workshops. September 2006, pp. 55-59, IEEE Computer Society.

[6] T. Yamakami, "An exploratory analysis on user behavior regularity in the mobile internet," in Proceedings of KES2006, Part III. October 2006, vol. LNAI 4253, pp. 143-149, Springer Verlag.

[7] 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. Gaber and P. Yu, "A framework for resource-aware knowledge discovery in data streams: a holistic approach with its application to clustering," in Proceedings of SAC'06. April 2006, pp. 649-656, ACM Press.

[10] T. Yamakami, "A time slot count in window method suitable for long-term regularity-based user classification for mobile internet," in MUE 2008. April 2008, pp. 25-29, IEEE Computer Society Press.

[11] T. Yamakami, "A long interval method to identify regular monthly mobile internet users," in AINA2008 Workshops/Symposium (WAMIS 2008). March 2008, pp. 1625-1630, IEEE Computer Society Press.

[12] T. Kamada, "Compact HTML for small information appliances," W3C Note, 09-Feb-1998, Available at: http://www.w3.org/TR/1998/NOTE-compactHTML-19980209, February 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.

Authors

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.