

Customer Analysis of Monthly-Charged Mobile Content Aiming at Prolonging Subscription Period

Yuji Shono, Yohei Takada, Norihisa Komoda
Graduate School of Information
Science and Technology, Osaka University
2-1, Yamada-oka, Suita 565-0871, Japan
Email: {yuji, yohe, komoda}@ist.osaka-u.ac.jp

Ayako Hiramatsu
Department of Information System Engineering
Osaka Sangyo University
3-1-1, Nakagaito, Daitou, Osaka 574-8530
Japan
Email: ayako@ise.osaka-sandai.ac.jp

Hiroaki Oiso
Codetoys K. K.
2-6-8 Nishitenma, Kita-ku, Osaka 530-0047
Japan
Email: oiso@codetoys.com

Kiyoyuki Fukaya
Faculty of Business Administration
St. Andrew's University
1-1, Manabino, Izumi, Osaka 594-1198
Japan
Email: fukaya@andrew.ac.jp

Abstract—Retention of customers is a critical challenge for mobile content providers. This paper presents a customer analysis result aiming at prolonging users' subscription. Based on users' score information, the access log, and so on, customers who are likely to unsubscribe in near future are identified. Regularity is discovered from the unsubscribed users' data using the C4.5 algorithm and a module library is prepared to easily formulate the character sets for the prediction from the users' access log. By using various character sets, the prediction of the users' subscription period was tested by using 6,000 real users' data. We predicted whether a user would unsubscribe within two months, with results showing 72% recall and 65% precision.

I. INTRODUCTION

The mobile content market in Japan has been expanding rapidly since 1998. In the subscription-type mobile content business, the users actually pay not only a small monthly content fee but also a large packet charge. However, the content provider can receive only a monthly content fee, because the carriers only collect the monthly content fee as an allocation. Actually the share of the monthly content fee constituting the total user payment is extremely low. Therefore, for the content providers, the increase in game times played by each user is not concerned with content providers' profit, leading to the necessity for content providers to carry out effective marketing to attract more users and to prolong the subscription period per user [1].

The target mobile content of this research is a quiz game which is provided through three domestic carriers in Japan. In this paper, we predict that a user will unsubscribe in the near future. Though the analysis is similar with the Churn Analysis Model [2][3][4][5][6], which deals with unsubscription behavior of the carriers' customers, there are some points of difference. Churn Analysis Model only uses the access log of users. In contrast, the analysis of game content customers uses not only access log but also other data (e.g. score, stage, and so on).

In this paper, we predict customers' unsubscription by referring to basic user information, score information, and the access log using a data mining technique. The most critical point is to extract effective character sets. However, since a 15-month access log contains an enormous quantity of data whose size is several hundreds M bytes, it is no simple task to extract adequate characteristic values by trial and error. Moreover, effective characteristic values depend on changes of the game rules or the communication environment. For flexible extraction of characteristic values, this paper presents a module library.

II. MOBILE CONTENTS

A. Outline of the Game

- Game Provision
The game content is provided in the form of a quiz. The game is provided through three mobile carriers, and the same game rules are applied among them, although there are some minor differences with other factors (e.g.: display layout).
- Subscription
When users subscribe to the content, they register nicknames, mail addresses, age, and sex. It seems that some users register untrue information regarding their age and sex because there is no disadvantage in winning a game by doing so.
- The Game Rules
One game comprises a maximum of 15 questions. All questions are answered by choosing one correct answer from four choices.
If the first question is answered correctly, the player receives 10,000 points, and the points double every time the player gets a correct answer. It seems that a lot of players will be able to score full points because it is easy to cheat through a mobile phone. Consequently we count

not only the points but also the time required to answer the question into the final score, which is calculated at the end of each month. The best score in all games in each month is set as the user's score, and the user's score is the criterion for ranking. In this content, four stages of ranking - the first stage (the lowest stage) through the fourth stage (the highest stage) - are prepared. The ranking is announced at the end of each month and only the top 25% of players in each stage can move up to a higher stage. If users unsubscribe, all information till the moment is lost. Therefore, when the users resubscribe, they must start from the first stage.

Besides the normal quiz, "fastest finger first" is held once a week, a contest open to all players. The top 10 users among them win the right to move to a higher stage the next month. By taking this quiz, users can move to a higher stage with only a tiny amount of packet charge and effort.

- The Maximum Number of Games

At the beginning of service, the maximum number of games per day was limited to five. Because of users' requests, however, this rule was changed in the summer of 2003. This means the maximum number of games per month is now set up to 150 instead of having the daily limit. Moreover, any surplus games can be carried over to the next month. This means if a user plays only 120 times in a certain month, he/she can play 180 times in the following month.

- Game Fee

The content fee is JPY 180 (about 1.3 Euro) per month, while the packet charge is about JPY 4 per quiz (= 1 page). For example, if the average number of quizzes in a game is eight, ten pages in total, which includes two pages before and after the game, will be forwarded every game. Therefore, the packet charge amounts to about JPY 40 (about 0.3 Euro). If a player plays 150 times in a month, it will thus cost JPY 6,000 (about 45 Euro) per month (JPY 6,180 including the content fee). In the case of the newly developed JAVA application version, the packet charge is JPY 20 per download of the 90-question quiz pack and JPY 200 to 400 to download the application.

- Incentive Provision

To prolong more users' subscriptions, a prize is provided to the five leading scorers who are at the highest stage.

B. Users' Character

- Users' Sequential Months

The time users subscribe to the game is measured in sequential months, with Fig. 1 showing the rate of users who subscribed in a certain month of 2003. The horizontal axis is in sequential months, while the vertical axis is the rate of users under subscription. As shown in Fig. 1, by the end of the 0th or the 1st month, most users of this content unsubscribe. Therefore, prediction accuracy for users whose number of sequential months is short should

be prioritized to prevent unsubscribing. Preventing the unsubscribing of users who cancel their subscription in one or two months is the best way to improve provider profit.

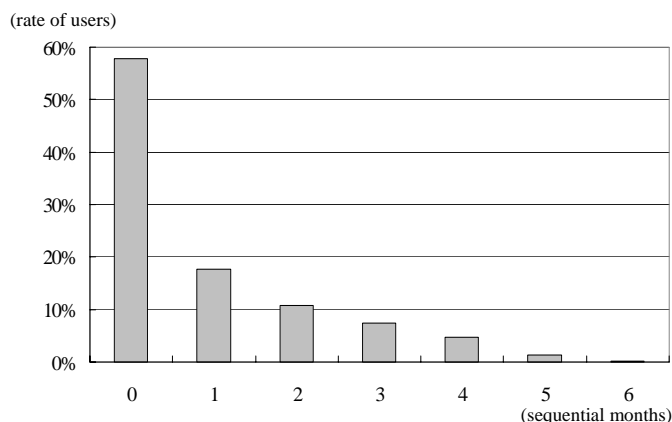


Fig. 1. Sequential months of users who subscribed in a certain month of 2003

- Users' Game Time

Each user's game time is quite different, depending on how much free time they have. The distribution of games played per hour is shown in Fig. 2. Users with jobs usually tend to play games from 12 a.m. to 1 p.m., which is lunch hour, and from 11 p.m. or later. However, night people and housewives are not applicable to this tendency.

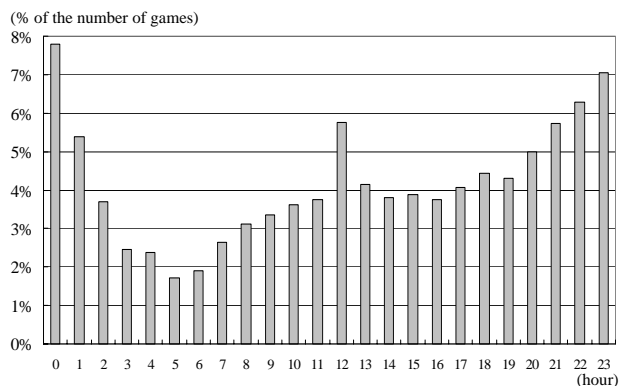


Fig. 2. Time of day when users play

III. UNSUBSCRIPTION AND DISSATISFACTION PREDICTION SYSTEM

A. Outline of the System

We are developing an unsubscribing and dissatisfaction prediction system for the content described in Chapter II. Fig. 3 shows an outline of customer analysis. By data mining from the access log and score information of users who have already unsubscribed, it is possible to predict the tendency of users who are going to unsubscribe. Based on the extracted tendency, a rule is derived. By applying the rule to the current users, it is possible to predict when they will unsubscribe, and this

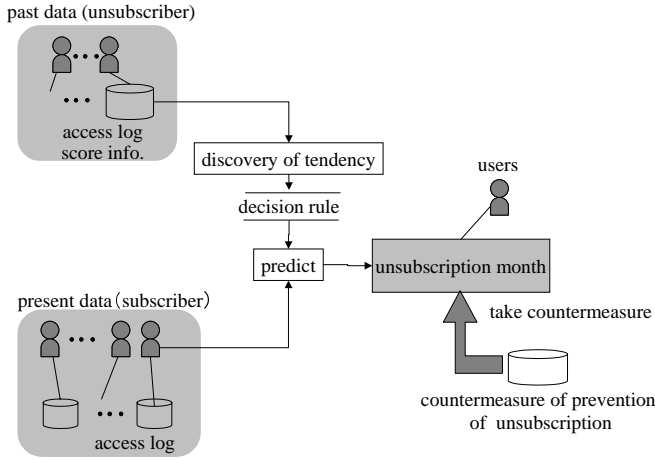


Fig. 3. Outline of customer analysis

prediction operation is applied to each user. If this prediction can be accurately accomplished, the content provider can take effective countermeasures against users who are going to unsubscribe.

Fig. 4 shows the system constitution for customer analysis. By referring to basic user information, the access log, and

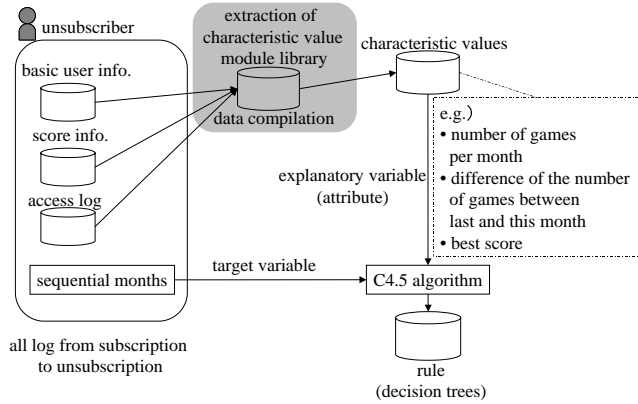


Fig. 4. System configuration

score information of unsubscribed users, a rule is generated to explain how long users will continue subscription. First, characteristic values, which are data of users' characters derived from the access log and the score information, are calculated and selected. A decision tree is then generated by a learning algorithm from the selected characteristic values. Because changes to the game rules and the communication environment greatly affect users' behavior, the characteristic values that can explain users' character often need to be changed.

Selection of the characteristic values is accomplished by trial and error until a good rule is generated, but a mechanism that can easily obtain the desired characteristic values is necessary. To deal with this requirement, a module library is prepared written in Perl. The module library is a collection of counting and aggregation programs.

Next, the characteristic values are selected in the same way by referring to each subscriber's access log and score information. By applying a decision tree to the characteristic values, we can predict how long users will stay subscribed. Thus, subscription duration can be predicted for each user. The number of months until unsubscribing is independently predicted. In other words, a decision tree is also generated for the prediction of unsubscription.

B. Access Log and Score Information

In this section, we explain the data used in the analysis. There is basic user information (user ID, age, etc.), which is registered at the beginning of subscription. Score information, which is calculated at the end of each month, consists of the following data. The quantity of score information is as many as the number of subscribing users who reach the end of the month, and the amount of data held by the carrier with the largest number of users is about 100,000 records in a database for 15 months. The score information includes following items.

- Current stage
- The number of each user's plays per month
- The best score per month
- Score rank
- Possibility of movement to a higher stage next month
- The number of users who answer all questions correctly

In addition to the data explained above, as shown in Table I, every time users access the Web page (= every time users play a game), user ID, date, score, and game time are recorded to an access log, which is a text file. The access log for 15 months has about 5,000,000 lines and the size is 760 Mbyte. It includes all users' access information in sequential order of access time. Therefore, to find meaningful values, a module library is prepared to compile data for flexible extraction of characteristic values. After extracting characteristic values of each user from the access log, they are recorded to a database.

TABLE I
ACCESS LOG

User ID	Date	Score	Game time (ms)
0728	2002/11/21 20:59:17	10,000,000	281,049
4393	2002/11/21 21:08:53	100,000	46,373
3255	2002/11/21 21:15:44	100,000	32,499
0728	2002/11/21 21:16:33	0	19,224
2041	2002/11/21 21:16:52	2,500,000	192,636
4393	2002/11/21 21:18:08	1,000,000	114,709
9887	2002/11/21 21:21:40	100,000	24,672
0728	2002/11/21 21:24:32	100,000	59,510
:	:	:	:

C. Module Library

Four functions of the module library prepared in this analysis are explained as follows. Fig. 5 shows that by using the four functions, the access log of each user is extracted and characteristic values are extracted.

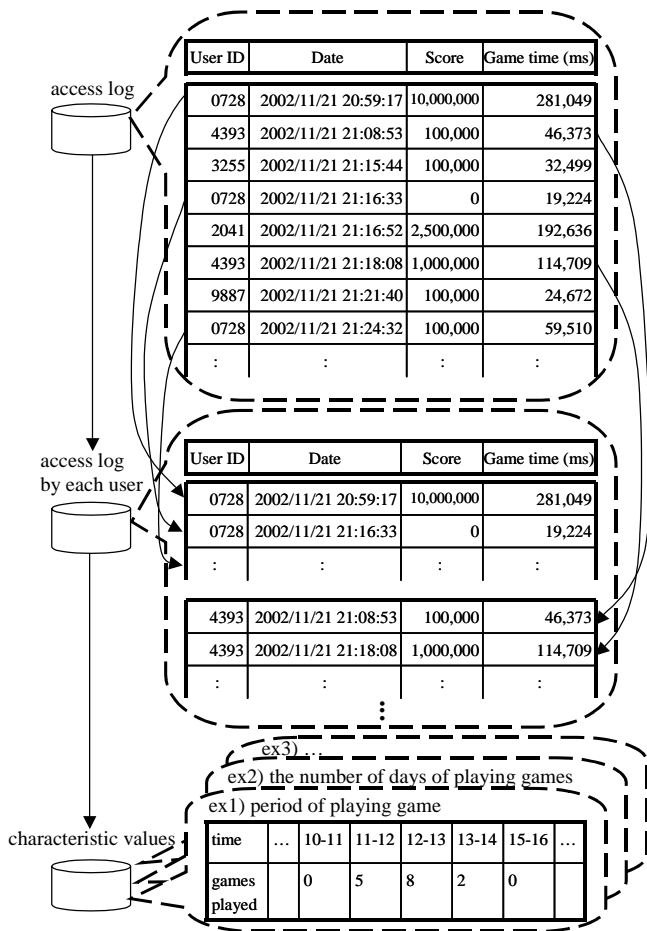


Fig. 5. Extraction from characteristic values using the module library

- Four basic arithmetic operations
The four basic arithmetic operations are addition, subtraction, multiplication, and division. For example, by using subtraction, we can determine the difference between the number of games played last month and this month.
- Converting the access log of all users to the access log of each user
The access log is not actually information on each user. Instead, it means that all users' logs are included in the access log. Therefore, an enormous quantity of access log of all users can be classified for each user. From each user's access log, we can extract several kinds of information such as those below.
 - The number of games per hour
 - The number of games early/late in the month
 - The number of days on which the user plays the game at least once
- Classification of users based on users' game time
Users' game time is particularly important because this highlights users' characteristics. Because users' game time differs greatly according to each user's job or for other reasons (written in 2.B), the number of games

played in every hour is counted. Thus by using the module library, it is possible to define characteristic values. For example, users who play many games at from 11:00 to 13:00 are considered as the noon-type users.

- Conversion of continuous values to discrete values
Continuous values such as the average number of games can be converted to discrete values by setting up thresholds. By changing a discrete value, users' characteristics are treated roughly. For example, a discrete value, "game frequency," is converted from the average number of games on a day X by using two thresholds M and N.
 - If $X < M$, "game frequency" is set to "low"
 - If $M \leq X \leq N$, "game frequency" is set to "medium"
 - If $X > N$, "game frequency" is set to "high"

Also, thresholds can be set up in each appointed period for change to a game rule and so on. For example, there are two periods: before/after a change in the limit of the maximum number of games. After the change to the game rules, many users play much more than before, thus the daily average number of games of most users increases. To respond to the change in the average number of games on a day, the two thresholds are modified to M' and N' after the change to the game rules.

It is easy to modify the definition of the characteristic values by rewriting the variables defined at the beginning of the program files.

D. Extraction of Characteristic Values

It is important to carefully select characteristic values for data. Fig. 6 shows the procedure to extract characteristic values with the learning algorithm from real data. Among provided users' data, basic user information and score information have already been registered in a database. By compiling and referring to the access log, several other characteristic values can be derived. By joining three databases (basic user information, score information, and each user's information from the access log) with SQL, we can obtain a characteristic values database to be used for the learning algorithm can be gotten.

E. Prediction

The C4.5 algorithm [7][8] is used as a learning algorithm. This algorithm outputs a decision tree. Data classified by sequential months is used as learning data for generating a decision tree, because the quantity of data is different according to the time. To evaluate the prediction, recall and precision are used. The recall shows the extraction rate; therefore a large number is better, while the precision indicates the noise rate. Here a larger number means less noise. They are defined by the following formulas.

$$\text{Recall} = \frac{\text{the amount of predicted relevant information}}{\text{the total amount of relevant information}}$$

$$\text{Precision} = \frac{\text{the amount of predicted relevant information}}{\text{the total amount of predicted information}}$$

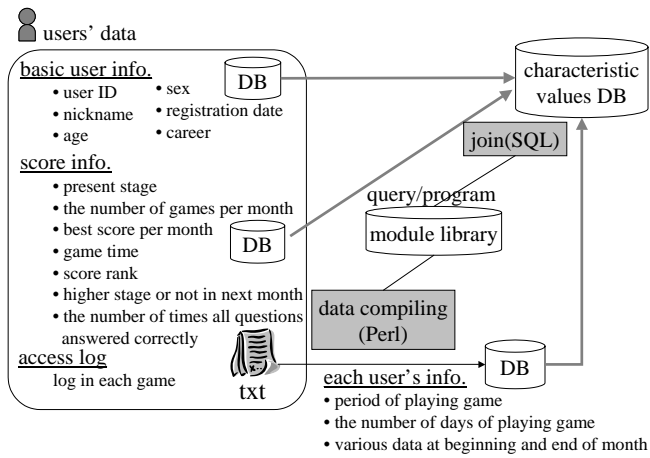


Fig. 6. Procedure to extract characteristic values for the learning algorithm

IV. EXPERIMENT FOR UNSUBSCRIPTION PREDICTION

A. Character Sets

In this experiment, we first prepared the four characteristic values sets as shown in Fig. 7. Character set (A) comprises basic information. Although character set (B) is almost the same as character set (A), character set (B) includes “users’ game time” but does not include “age” and “sex,” which have less data reliability. Character set (C) is basically the same as character set (B). However, character set (C) includes continuous characteristic values. Character set (D) is basically same as character set (C). However, character set (D) includes discrete values that are converted from continuous values of character set (C), based on the thresholds.

B. Prediction Results

We examined the prediction accuracy of the four character sets. A total of 6,000 unsubscribed users in seven months were extracted at random and they divided into two groups. One was for learning data to generate a decision tree, while the other is for test data to examine prediction accuracy.

Fig. 8 shows a comparison of recall of unsubscrition prediction for whether users unsubscribe within two months. This prediction was performed by using every character set. The horizontal axis represents how long users are under subscription (sequential months). The vertical axis is recall. Prediction accuracy was changed by changing characteristic values. As shown in Fig. 8, in the case of character set (D) and one sequential month, the recall is 71.6% which is better than the result for other character sets. For the case of character set (D) and two or three sequential months, the recalls are a little better than those of other character sets. In this content, since more than half the users unsubscribe by one or two months after subscribing, high prediction accuracy for short sequential months is important.

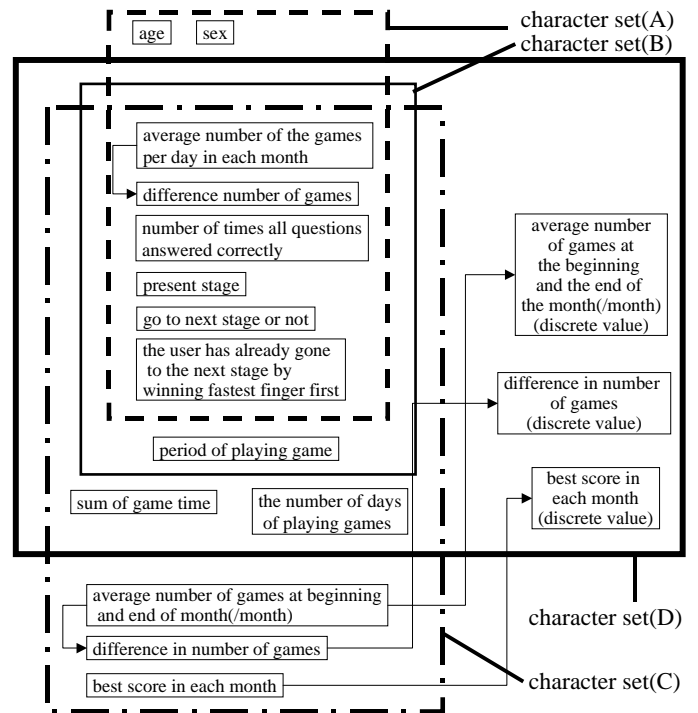


Fig. 7. Preparing character sets

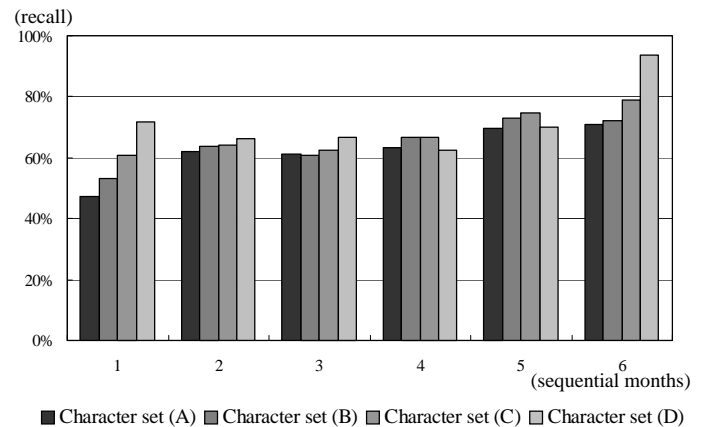


Fig. 8. Comparison of recall for every character set

C. Unsubscrition Prediction Accuracy

In Table II shows in detail prediction results in the case of using character set (D). “Classification result” in the table shows the result for prediction of users’ unsubscrition. Classification result (a) is for users who are predicted to unsubscribe within two months after subscribing. Classification result (b) is for users who are predicted to unsubscribe after two months.

The evaluation only defines whether the prediction of users who unsubscribe within two months can actually be accomplished. Recall and precision in the case where users are selected at random is assumed to be same as rate for the test data. In this case, we have confirmed a recall improvement

TABLE II
RECALL AND PRECISION (UNSUBSCRIPTION)

Sequential Months	Objective Variable	Learning Data	Test Data			Prediction Result			
			Total Data	Unsubscriber	Rate (%)	Classification result		Recall (%)	Precision (%)
						(a)	(b)		
1	users who unsubscribe within 2 months	1633	3000	1687	56.2	1208	479	71.6	64.6
	users who unsubscribe after 2 months	1367		1313	43.8	661	652	49.7	57.6
2	users who unsubscribe within 2 months	1071	2000	1072	53.6	710	362	66.2	63.6
	users who unsubscribe after 2 months	929		928	46.4	407	521	56.1	59.0
3	users who unsubscribe within 2 months	798	1500	817	54.5	545	272	66.7	60.6
	users who unsubscribe after 2 months	702		683	45.5	355	328	48.0	54.7
4	users who unsubscribe within 2 months	533	1000	575	57.5	360	215	62.6	63.3
	users who unsubscribe after 2 months	467		425	42.5	209	216	50.8	50.1
5	users who unsubscribe within 2 months	418	700	432	61.7	303	129	70.1	63.1
	users who unsubscribe after 2 months	282		268	38.3	177	91	34.0	41.4
6	users who unsubscribe within 2 months	296	450	289	64.2	271	18	93.8	66.6
	users who unsubscribe after 2 months	154		161	35.8	136	25	15.5	58.1

of 10 to 15% and a precision improvement of 5 to 10% is confirmed.

D. Response to Change of a Game Rule

Here, the effectiveness of changing a characteristic value was examined. In the summer of 2003, a game rule was changed. Before that, the maximum number of games in a day was five, but following that, the daily limit was scrapped and the maximum number of games per month was established as 150.

To decide a discrete characteristic value, "game frequency" from the average number of games, six thresholds are used. The best two thresholds values tuned for before and after the rule change are shown in Table III.

TABLE III
THRESHOLDS TUNED FOR BEFORE AND AFTER THE RULE CHANGE

Game Frequency Value	Thresholds tuned for before and after the rule change	
	before	after
L1	less than 0.5	less than 0.5
L2	less than 1	less than 1
L3	less than 2	less than 3
L4	less than 3	less than 6
L5	less than 4	less than 9
L6	more than 4	more than 9

As a result of changing a characteristic value definition, recall is improved by 7.8% in case of character set (D). It was confirmed that changing characteristic values to respond a change of a game rule was effective.

V. CONCLUSION

By preparing multiple character sets, it was possible to predict unsubscription of the game content to approximately 72% recall and 65 % precision. However, the prediction accuracy needs to be improved for practical use. To improve the prediction accuracy, first, it is advisable to improve the cutback of branches of a decision tree from the C4.5 algorithm. Improving the cutback may lead to improved prediction

accuracy. Second, after classifying such users in advance, it is effective to generate decision trees for each classified group.

Users' unsubscription is caused by various reasons. Dissatisfaction is one reason for unsubscribing, thus, it is necessary to take suitable action to reduce the number of dissatisfied users is desired. To do this, in addition to unsubscription prediction, a dissatisfaction-prediction analyzing a questionnaire at the time of unsubscription is necessary. One piece of relevant research in our research group is "A Method for Atypical Opinion Extraction from Answers in Open-ended Questions"[9].

The communication environment is rapidly changing. For example, a flat packet fee system will be introduced, and those changes will affect users' behavior. Therefore, for the provider to use the proposed analysis method, a man-machine system supporting the extraction of characteristic values is needed.

REFERENCES

- [1] H. Oiso and N. Komoda: "Access Analysis of Monthly-Charged Mobile Content Provision," in *Proc. of Future Business Technology Conf. (FUBUTEC'2004)*, pp. 76-80 (2004).
- [2] R. Mattison: *Data Warehousing and Data Mining for Telecommunications*, Artech House (1997).
- [3] S. Rosset, E. Neumann, U. Eick, N. Vatnik, and Y. Idan: "Customer Lifetime Value Modeling and Its Use for Customer Retention Planning," in *Proc. of the 8th ACM SIGKDD Inf. Conf. on Knowledge Discovery and Data Mining*, pp. 332-340 (2002).
- [4] M. J. A. Berry, G. Linoff: "Mastering Data Mining (case example edition) -art and science of CRM-," Wiley & Sons Inc. (2000).
- [5] S. Rosset, E. Neumann, U. Eick, and N. Vatnik: "Lifetime Value Models for Decision Support," *Data Mining and Knowledge Discovery Journal*, Vol. 7, pp. 321-339 (2003).
- [6] D. R. Mani, J. Drew, A. Betz, and P. Datta: "Statistics and data mining techniques for lifetime value modeling," in *Proc. of the fifth ACM SIGKDD Inf. Conf. on Knowledge Discovery and Data Mining*, pp. 94-103 (1999).
- [7] J. R. Quinlan: *C4.5: Programmers For Machine Learning*, Morgan Kaufmann (1993).
- [8] "Ross Quinlan - AI Group, CSE," <http://www.cse.unsw.edu.au/quinlan/>
- [9] A. Hiramatsu, S. Tamura, H. Oiso, and N. Komoda: "A Method for Atypical Opinion Extraction from Answers in Open-ended Questions," in *Proc. of IEEE International Conference on Computational Cybernetics (ICCC 2004)*, (2004).