Analysis Using Popularity Awareness Index, Recency Index and Purchase Diversity in Group Buying

Yasuyuki Shirai[1], Hiroyuki Morita[2], Masakazu Nakamoto[1], and Satoshi Oyama[3]

[1] JST-ERATO Minato Discrete Structure Manipulation System Project, Hokkaido University, Sapporo, Japan
{shirai,nakamoto}@erato.ist.hokudai.ac.jp
[2] College of Sustainable System Sciences, Osaka Prefecture University, Osaka, Japan
morita@eco.osakafu-u.ac.jp
[3] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan
oyama@ist.hokudai.ac.jp

Abstract. We propose new metrics for customers' purchasing behaviors in a group buying coupon website, based on HITS algorithms and information entropy: that is, popularity awareness index, recency index, and purchase diversity. These indices are used to classify customers and predict future behaviors. This paper includes definitions of these new indices to be used in real group buying websites. In these websites, adequate characteristics for customers are strongly required and are critical for marketing purpose. We will also provide some experimental results on real data set, including customer segmentation used in future marketing planning.

Keywords: group buying, HITS algorithm, entropy.

1 Introduction

Various types of e-commerce have grown steadily around the world. In Japan, the domestic business-to-consumer (B-to-C) e-commerce market size has reached 9.5 trillion yen in 2012 [1]. In addition, the business models of e-commerce transactions have diversified accordingly, with the emergence of group buying coupon sites. In this model, retailers exhibit a coupon in the group buying coupon site restricting purchase time limit, conditions of purchase, and so on. In addition, the retailers set a limit to the minimum quota of transactions. For instance, if total volume of coupons purchased by customers does not exceed the minimum volume, the deal will not be established. Although setting the minimum volume does not seem to facilitate sales, customers who want to purchase the coupons would refer their friends to sign up for the coupon in order to propagate enough quota to establish the transaction within the purchase time limit. This

P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 100–114, 2014.

framework is referred as flash marketing. GROUPON and PONPARE are two big group buying coupon websites in Japan, with coupon sales averaging at 1.5 billion yen every month [2]. However, the market share changes rapidly, thus every site has a chance to lead the market. In order to outperform against other rival companies, identification of proper customer segmentation and adequate implementation of Customer Relationship Management strategy are necessary.

In this paper, we analyze the web log data and related data provided from a group buying coupon site in Japan[1]. The data consists of one year of transactions from July 2011 to June 2012. Although the behavior and purchased history which can be used for analysis are sampled coarsely to preserve privacy, personal attributes including age and gender are distributed. Information about the coupon, namely, the name of the coupon, fixed price, purchased price, discount rate, genre of the coupon are also available. From our preliminary analysis of the data, we have noted the following attention points.

1. The stakeholders of this site are the site managing company, retailers, and purchasers. And the retailers exhibit only one coupon at one time without repeating the exhibition.
2. The data is coarsely-sampled as we mentioned. Even if we aggregate total number of coupons purchased, we will not be able to grasp the entire behavior of the customers.
3. Most of the coupon discount rates are more than 50% (high level of discount). Moreover, there are several coupons whose discount rates are 100%.

From the above observations, it is difficult to focus on both the price and the continuous selling in our analysis. Instead we think that it is more important to analyze the relationship between customers and the coupons. Purchasing behaviors can be classified into various types. Some purchasers may only be interested in popular coupons among others purchasers, and some purchasers may be interested in unique coupons.

We think that focusing on the popularity of the coupon with such purchasing characteristics will be very insightful. Thus, we have defined coupon popularity awareness index for analyzing the customer's behavior based on the index. To do this, we propose using Hypertext Induced Topic Selection algorithm (HITS) [5] to define the index. HITS calculates hub store and authority score between websites from web log data and derives the importance of the website of the search engine. Although the relationship among websites can be expressed by general graph, the relationship between customers and coupons can be expressed by bipartite graph. We apply HITS algorithm to such bipartite graph and calculate the hub and authority score in the same way as usual HITS. In this paper, we shall define the scores as coupon popularity index and popularity awareness index, respectively.

Additionally, we focus on the customers' response when making purchases. We further defined the customers' response to the release of coupons on the

[1] The data is provided by Joint Association Study Group of Management Science at Data Analysis Competition in 2012.

website : for example, some customers purchase coupons immediately after the release, while some customers purchase coupons after observing other customers' behaviors. Figure 1 illustrates the relationship between the elapsed time after the release and the sales volume[2]. We discovered that the highest volume of

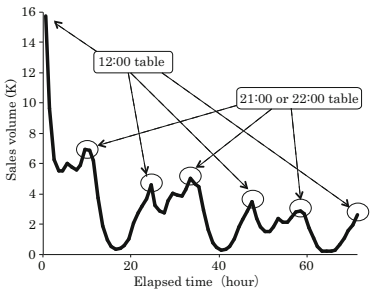

Fig. 1. Transition of purchasing from the coupon release

purchase takes place during the first two hours from the release. The cycle is repeated in 24 hour intervals at a decreasing rate. In flash marketing, purchase time limit is short, so we can use the difference between the responses as the basis of customer segmentation defined as "recency index" in later section.

Finally, the variety of coupons purchased also provides an important perspective. Customers who purchase coupon in only one genre easily lose their interest about the genre, so they may stop purchasing coupons from the site. However, customers who purchase coupons in various genres are more likely to retain their interest, even if they lose their interest in only one genre. To reflect this viewpoint, we apply the concept of information entropy, which is referred to as "purchase diversity".

The following sections define customer's segmentation using three indices mentioned above with corresponding analyses.

2 Indices for Customers and Coupons

In this section, we will first describe the data set, and then show the definition of each index and some preliminary results using those indices.

2.1 Overview of the Data

The data are coarsely-sampled and have some restrictions as mentioned earlier. Since fragmented data may create noise in our analysis, we have filtered this data set so that each user has at least two purchased coupons, and each coupon has at least two purchasers. Finally 7573 customers and 6646 coupons are extracted.

[2] New release time of this web site is almost noon.

Table 1 shows the distribution of the number of customer purchases and the number of coupon sales.

Coupon genres included in the original data set are designed for the real operation in the group buying website. Hence, we have rearranged coupon genres for the purpose of our analysis as follows : 'restaurant', 'relaxation', 'health & beauty', 'grocery & gourmet food', 'travel' and 'cosmetics'.

Table 1. The Number of Purchases by Customers and the Number of Coupon Sales

Num. of purchases (N)	Num. of customers
$2 \leq N < 5$	2663
$5 \leq N < 10$	2439
$10 \leq N < 15$	1272
$15 \leq N < 20$	573
$20 \leq N < 25$	283
$25 \leq N < 30$	160
$30 \leq N$	183
total	7573

Num. of sales (M)	Num. of coupons
$2 \leq M < 5$	3131
$5 \leq M < 10$	1719
$10 \leq M < 15$	731
$15 \leq M < 20$	344
$20 \leq M < 25$	209
$25 \leq M < 30$	109
$30 \leq M$	403
total	6646

2.2 Popularity Awareness Index (PAI) and Coupon Popularity Index (CPI)

HITS algorithm (Hyperlink-Induced Topic Search)[5,6] and PageRank algorithm [4] are well known web page ranking methods. The idea behind HITS algorithm is based on the concept of hubs and authorities in the Internet hyperlink structure. That is, a good authority page is linked by many good hub pages, and a good hub page has links to many good authority pages. HITS algorithm can rank each web page as hub site and authority site, based on the recursive definition.

We apply the HITS algorithm to explain relations between customers and coupons. In shopping behaviors, the concept of 'goodness' is not as obvious as Web pages. So we define 'popular coupons' as coupons purchased by customers who are conscious of popularity, and these customers who purchase popular coupons as 'popularity-aware customers'. Based on these recursive concepts, we can rank the 'popularity awareness' and 'coupon popularity', which are called *popularity awareness index* (PAI) and *coupon popularity index* (CPI), respectively. The details of the ranking methods are described as follows.

Let the PAI for customer i be u_i, and the CPI for coupon j be m_j. We consider the vector of PAIs : $\boldsymbol{u} = (u_1, u_2, \ldots, u_{n_u})^{\mathrm{T}}$, and the vector of CPIs : $\boldsymbol{m} = (m_1, m_2, \ldots, m_{n_m})^{\mathrm{T}}$ where n_u shows the number of customers and n_m shows the number of coupons. We can compute $\boldsymbol{u}$ and $\boldsymbol{m}$ by assigning the initial value $\boldsymbol{m}^{(0)}$ and calculating iteratively, where N denotes the number of repeat count, and $\boldsymbol{u}^{(N)} = (u_1^{(N)}, u_2^{(N)}, \ldots, u_{n_u}^{(N)})^{\mathrm{T}}$, $\boldsymbol{m}^{(N)} = (m_1^{(N)}, m_2^{(N)}, \ldots, m_{n_m}^{(N)})^{\mathrm{T}}$. In the following equations, $i \to j$ represents that customer i purchased coupon j.

$$\overline{u}_i^{(N+1)} = \sum_{j:i \to j} m_j^{(N)} \quad (i = 1, \ldots, n_u)$$

$$\overline{m}_j^{(N+1)} = \sum_{i:i\to j} u_i^{(N+1)} \quad (j = 1, \ldots, n_m)$$
$$\boldsymbol{u}^{(N+1)} = \overline{\boldsymbol{u}}^{(N+1)} / \|\overline{\boldsymbol{u}}^{(N+1)}\|_2$$
$$\boldsymbol{m}^{(N+1)} = \overline{\boldsymbol{m}}^{(N+1)} / \|\overline{\boldsymbol{m}}^{(N+1)}\|_2$$

$\|\cdot\|_2$ represents L_2 norm, that is, each $\boldsymbol{u}^{(N+1)}$ is normalized so that square sum of each element is 1. Now let $a_{ij} (1 \le i \le n_u, 1 \le j \le n_m)$ be the purchase flag (0 or 1) by customer i for coupon j, and let

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n_m} \\ a_{21} & a_{22} & \ldots & a_{2n_m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_u 1} & a_{n_u 2} & \ldots & a_{n_u n_m} \end{pmatrix},$$

then we can get

$$\boldsymbol{u}^{(N+1)} = \boldsymbol{A}\boldsymbol{m}^{(N)}, \quad \boldsymbol{m}^{(N+1)} = \boldsymbol{A}^{\mathbf{T}}\boldsymbol{u}^{(N+1)}.$$

From the above mutual recursive equations, we can get

$$\boldsymbol{u}^{(N+1)} = \boldsymbol{A}\boldsymbol{A}^{\mathbf{T}}\boldsymbol{u}^{(N)}, \quad \boldsymbol{m}^{(N+1)} = \boldsymbol{A}^{\mathbf{T}}\boldsymbol{A}\boldsymbol{m}^{(N)}.$$

It is well known that $\boldsymbol{u}^{(N)}$ and $\boldsymbol{m}^{(N)}$ converge to the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathbf{T}}$ and $\boldsymbol{A}^{\mathbf{T}}\boldsymbol{A}$ respectively. These normalized eigenvectors are shown by $\boldsymbol{u}, \boldsymbol{m}$. We calculated $\boldsymbol{u}, \boldsymbol{m}$ for 7573 customers and 6646 coupons based on the above framework. Moreover, we prepared PAIs for the first 6 months in order to observe the situation for the first 6 months. There are 3936 customers who have made purchases in July and August, 2011, therefore we calculated the indices based on the purchase behaviors of these customers until the end of December 2011.

We define the rank of PAIs sorted by descending order as 'popularity awareness index rank (PAI rank)'. The top-ranked customer is more aware of and sensitive to coupon popularity when compared to other customers.

2.3 Recency Index (RI)

We define the *recency index* (RI), P_i for customer i, which represents customer's purchase response. Let I_i be a set of purchased coupons by customer i, $r_k(i)$ be the order of purchase by customer i among all customers who purchased coupon $k (\in I_i)$, and M_k be the total number of purchasers for the coupon k. For example, suppose that there are 100 purchasers for a coupon k, and customer i is the 10th customer to purchase the coupon, $M_k = 100$, $r_k(i) = 10$.

We will define the recency index P_i (RI), as a normalized value of the sum of differences between the number of purchases which took place before user i and the number of purchases after user i :

$$P_i = \frac{1}{|I_i|} \sum_{k \in I_i} ((M_k - r_k(i)) - (r_k(i) - 1)),$$

In comparison, customers with negative RIs purchase coupons after other customers, while customers with positive RIs purchase coupons ahead of other customers.

As is the case with PAIs, in order to observe users' behavior during the first 6 months, we define 'RI for the first 6 months' based on the purchase behavior until December 2011, where customers have purchased in July or August 2011 (3936 persons). 'RI rank' also shows the rank of customers, sorted by descending order. The top-ranked customers in the RI make coupons purchases faster than other customers.

2.4 Purchase Diversity (PD)

Some users may only purchase coupons of a specific genre, whereas other users may purchase coupons in various genres. To quantify the diversity of purchases, we apply the concept of information entropy.

Purchase diversity (PD) can be defined using the purchase ratio for each genre as follows. Let i ($i = 1, \ldots, n$; where n is a number of customers) be a customer, and j ($j = 1, \ldots, m$; where m is a number of genres) be a genre. Thus the PD, B_i for customer i, can be defined as follows.

$$B_i = -\sum_j p_i^j \log_2 p_i^j$$

where p_i^j refers to the purchase ratio of genre j for the total purchase amount by user i. Table 2 shows the simple example of the number of purchases in each genre by each user and the PD for each user, which represents a degree of variety of purchased coupons by each customer.

Table 2. Example of PD (Purchase Diversity)

Customer	restaurant	travel	cosmetics	relaxation	PD
id1	3	2	1	2	1.906
id2	0	0	5	1	0.650
id3	10	0	0	0	0.000

In fact, if a user i only buy coupons of a specific genre, $B_i = 0$. On the other hand, if a user buy coupons in various genres, the value of PD will be high. The maximum value of PD is determined by the number of genres. In our case, if we adopt six genres, the maximum value of PD is $\log_2 6 \simeq 2.585$.

3 Experimental Results

3.1 Distribution of the Basic Indices

Popularity Awareness Index (PAI) Table 3 shows the top 10 customers with high PAIs, at which these top tier customers prefer popular coupons. For

Table 3. Top 10 Customers in PAIs

Sex	Age	PAI	Number of Purchase						Total Price (Yen)	
			restaurant	relaxation	health and beauty	grocery & gourmet food	travel	cosmetics	catalog price	cut-rate price
m	36	0.11348	2	0	0	0	1	8	81710	23025
f	36	0.09160	0	8	6	3	0	16	298125	76353
m	36	0.08995	19	0	0	1	1	7	124274	38772
f	55	0.08380	22	5	0	14	5	4	301307	131878
f	38	0.07688	1	1	0	2	1	6	115335	30422
f	52	0.07675	23	0	0	3	0	4	173451	54692
f	50	0.07412	25	1	0	0	2	4	246892	111273
f	34	0.07326	4	0	0	1	0	8	65947	22447
f	28	0.07277	22	5	2	5	3	3	412171	152693
f	42	0.07259	3	1	0	2	3	6	170490	45115

example, the top customer in this list purchased 1st, 8th, 20th, 47th, 618th coupons in CPIs.

Table 4 shows the comparison of average PAI ranks by gender and age bracket. It is noted that the PAIs of female customers are, in general, higher than male customers. This implies that middle-aged women have a higher tendency to purchase popular coupons.

Table 4. The Comparison of PAI Ranks (Average) by Gender and Age Bracket

Gender	20's	30's	40's	50's	60's
Female	4440	4071	3450	3195	3032
Male	4495	4066	3931	3695	3834

As for the prediction ability of future behaviors, we compared the behaviors for the last 6 months for group with high PAI against group with low PAI for the first 6 months (the two groups are divided based on average value). The average frequency (during the last 6 months) of high PAI group is 5.306, whereas the frequency of low PAI group is 3.201. This result shows that the PAIs are strong explanatory variables to predict future behaviors.

Coupon Popularity Index (CPI). Table 5 shows the top five coupons in CPI. The catalog price and the discount rate are shown in the table, however, there are no correlation in fact, between CPI and catalog price, or CPI and the discount rate. Thus it is concluded that the coupon popularity does not correspond to the price. It is observed that there are many cosmetics and restaurant coupons among the popular coupons. Finding out the characteristics or features for popular coupons can be the future research problem.

Table 5. Top 5 Coupons in CPIs

Genre	Item Name	Number of Sales	CPI	Discount (%)	Catalog Price (Yen)
cosmetics	Uruwoeet facial mask	463	0.584	78	7200
restaurant	Baumkuchen (Special Price)	353	0.176	50	1260
restaurant	Waffle Set (Special Price)	221	0.167	54	1100
travel	Pair Ticket with Breakfast	290	0.160	55	17000
cosmetics	White Moisture Cream	251	0.157	81	7980

Recency Index (RI). Figure 2 shows the distribution (number of customers) of RIs. RIs are distributed almost symmetrically around the center. As with PAI, we compared the behaviors during last 6 months, for group with high RI against the group with low PAI in the first 6 months (the two groups are divided based on average value). The average frequency of purchase for the high RI group during last 6 months is 4.444, whereas the frequency for low PAI group is 4.061. Although RIs are less effective explanatory variables than PAIs, RIs still contain informative features for future activities.

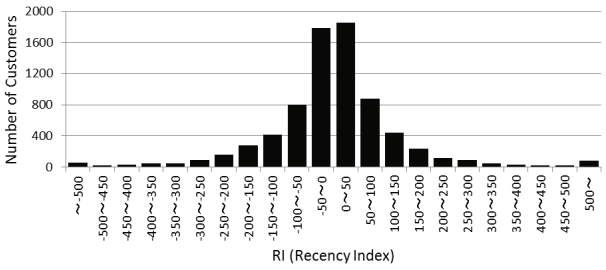

Fig. 2. Distribution of Recency Points

As shown previously, the relationship between PAIs/RIs for the first 6 months, and average purchase frequencies for the last 6 months is significant. Combining these variables enables the precise prediction of purchase frequency in the future. Table 6 shows the relationship between PAIs and RIs for the first 6 months, and average purchase frequency for the last 6 months. By combining these two parameters, the frequencies of the group with high PAIs and RIs is two times higher in the last 6 months compared to the group with lower PAIs and RIs.

Purchase Diversity (PD). Table 7 shows the comparison of numbers of purchases for last 6 months by increase/decrease of PD corresponding to the number of purchases for the first 6 months (N). Although the difference between decrease and increase is not significant when $10 < N$, there are clear differences in the

Table 6. PAIs and RIs for the First 6 Months and the Average Purchase Frequency for Last 6 Months

	RI (High)	RI (Low)
PAI (High)	5.531	5.069
PAI (Low)	3.304	3.103

case of $N \leq 5$ and $6 \leq N \leq 10$. In fact, in these two cases, the frequency of the increase group is more than 150% than the decrease group. From the results, we can conclude that purchase variety is considerably important if we expect continual activity for the last 6 months. Therefore it is recommended to promote a wide variety of coupons to the group with low purchases as marketing strategy. On the other hand, some customers in the high purchase group have solid base of coupons purchased, thus it is not necessary to promote other genre of coupons for these users.

Relationship Among PAIs, RIs and PDs. So far we have defined three kinds of parameters to characterize each customer's behavior, that is, popularity awareness index (PAI), recency index (RI) and purchase diversity (PD). Figure 3 shows the relationship among three indices, where each data point in the plot represents a customer, and "CC" is correlation coefficient. From this observation, we use these three indices as independent explanatory variables in the next subsection.

3.2 Analysis for Business Application

In this section, PAIs, RIs and PDs are analyzed from the business application viewpoint. First of all, we define active customers and dormant customers, and build a decision tree model to classify these two classes. Customers who purchased coupons in July or August 2011, with three times or more purchases made during the first 6 months are selected. Of the selected customers, customers who purchased more than three months in the last 6 months are defined as active customers (ACTIVE), and customers who did not purchase any coupon for the same period are defined as dormant customers (DORMANT). The number of active and dormant customers are 1700 and 153 respectively. The classification factors generated from the decision tree models are examined.

Table 7. Comparison of the Numbers of Purchase for Last 6 Months

	Number of Purchases (N) for First 6 Months		
	$N \leq 5$	$6 \leq N \leq 10$	$10 < N$
Decrease PD(6 month $\rightarrow$ 1 year)	2.017	4.014	8.273
Increase PD(6 month $\rightarrow$ 1 year)	3.303	6.149	9.265

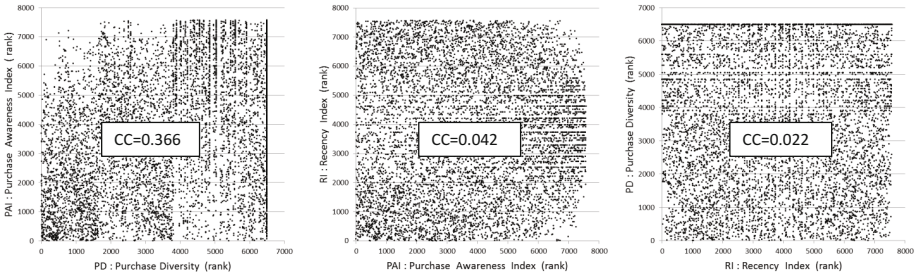

Fig. 3. Relation Among PAIs, RIs and PDs

Secondly, we are proposing a method to segment active customers with PAI rank, RI rank and PD rank. For each rank, customers are divided into two segments which are higher rank segment and lower rank segment by using the median of the ranking. Subsequently, we obtained eight customer segments. Of all segments, we focus on two segments with all higher rank or all lower rank, create classes in each segment, and build decision tree models used for classification problems.

Table 8. Explanatory Variables and Their Contents

Explanatory Variables	Contents
gender	male or female
age	The range is from 20 to 79.
NUM6	The number of total purchased coupons for the first 6 months.
restaurant	The number of purchased coupons for each genre for the first 6 months.
relaxation	
health_&_beauty	
grocery_&_gourmet_food	
cosmetics	
travel	
PAI6,PD6,and RI6	The rank of PAI, PD, and RI for the first 6 months, respectively.

In the following decision tree model, objective variables are based on the purchasing behaviors in the last 6 months. Thus the purchasing behaviors only for the first 6 months are used as explanatory variables. Table 8 denotes the explanatory variables from customer's characteristics and the purchasing behavior for the first 6 months. The decision tree models are constructed with the WEKA[3] J48 algorithm, and five folds cross validation method is used to test the model. In decision tree models shown in Figure 4, 6, 7, the description in the branch node denotes the branch condition, and the description in the leaf node denotes the branch condition, the predicted class, and the actual result for the prediction

[3] `http://www.cs.waikato.ac.nz/ml/weka/`

(the number of instances which match the rule, followed by how many of those instances are incorrectly classified, separated by "/").

Table 9 shows the evaluation index from the decision tree model for ACTIVE and DORMANT class. We can estimate the validity of class prediction using F-measure for both classes. Figure 4 shows the decision tree for active customers and dormant customers. According to the decision tree, the first division is decided by the number of coupons purchased from restaurant. The conditions of the key prediction rule (as shown in grey node (1)) of dormant customers are described as : restaurant coupons are purchased less than two times, travel coupons are purchased less than four times, and grocery and gourmet food coupons are purchased less than five times. On the other hand, the conditions of the key prediction rule of active customers are described as : restaurant coupons are purchased more than three and less than ten times, and cosmetics coupons are never purchased (as shown in grey node (2)), otherwise restaurant coupons are purchased more than nine times (as shown in grey node (3)). From the above observations, it can be noted that many of active customers purchase coupons in relation to restaurant, travel, and grocery and gourmet food categories simultaneously. Otherwise these customers are likely to purchase three or more times coupons in the restaurant category. From this result, we can conclude that customers who are satisfied with the coupon purchase in the restaurant, travel, or

Table 9. Evaluation Index of the Decision Tree Model for the Active Customer and the Dormant Customer

	Accuracy	Class	Precision	Recall	F-Measure	ROC Area
Training data	77.12%	ACTIVE	0.779	0.758	0.768	0.806
		DORMANT	0.764	0.784	0.774	0.806
Test data	63.07%	ACTIVE	0.645	0.582	0.612	0.669
		DORMANT	0.619	0.680	0.648	0.669

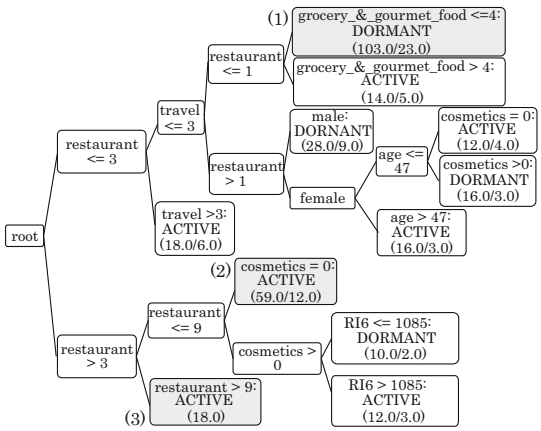

Fig. 4. Decision Tree for the Active Customer and the Dormant Customer

grocery and gourmet food categories, with repeated purchase of these coupons, are likely to become active customers.

Next, we will analyze active customers in detail. The PAI rank, RI rank and PD rank for the first 6 months and for 12 months are calculated respectively for active customers. Then, each index is divided into two segments, higher class and lower class after 6 months and after 12 months by using the median of ranking for each. The left side of Figure 5 shows eight segments after 6 months. Each box on the left side of the figure includes the values of PAI6, PD6 and RI6 (higher and lower), and the number of customers who belong to the segment. Of these segments, the segments which have all higher rank or all lower rank are called HIGH class and LOW class, respectively. Then two decision tree models were built as follows : The first one generates classification rules for the customers who belong to HIGH class. This model, which we call HIGH class model, classifies the group of higher class customers who remain in higher class (STAY), and group of lower class customers who have moved down to lower class (DOWN). The second one builds classification rules for customers who belong to the LOW class. This model, which we call LOW class model, aims to classify customers who remain in lower class (STAY) or who have moved up to higher class (UP).

Table 10. Evaluation Index of the Decision Tree Model for the Higher Class

	Accuracy	Class	Precision	Recall	F-Measure	ROC Area
Training data	81.11%	DOWN	0.680	0.836	0.750	0.870
		STAY	0.905	0.798	0.848	0.870
Test data	65.00%	DOWN	0.477	0.344	0.400	0.616
		STAY	0.706	0.807	0.753	0.616

Table 10 and Figure 6 illustrate the performance index and the decision tree of the HIGH class model, respectively. In this case, the number of customers in STAY class and DOWN class are 119 and 61, respectively. Table 10 shows the prediction for STAY class is more precise than the prediction for DOWN class for test data set.

The major rule antecedent that distinguishes STAY class holds the condition where the value of PD and PAI is higher (the PD6 and PAI6 is lower) as shown in grey node (1) of Figure 6, or when the number of relaxation coupons purchased are more than two as shown in grey node (2). On the other hand, the major rule antecedent that distinguishes DOWN class holds the condition when the number of relaxation coupon purchased is less than three, and the value of PD is low (PD6 is high), and NUM6 is less than 14 as shown in grey nodes (3) in Figure 6. From the above observations, high PD and high PAI are important classification factors to determine whether customers are likely to remain in higher class. On contrary, customers who have low diversity and purchase little coupons are likely to move down to a lower level segment. We can guess that diversity of coupons purchased and the popularity of coupons are related to user's interests in coupons. Thus, if the customer only purchases a few homogeneous coupons, it

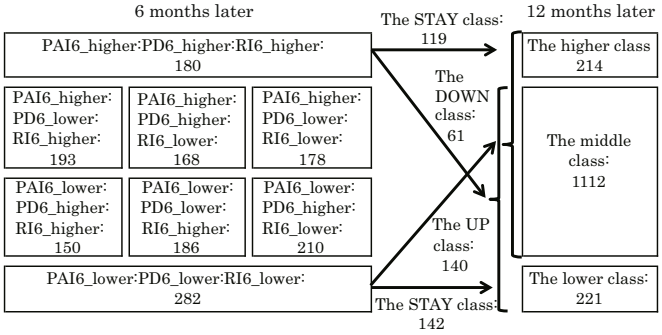

Fig. 5. Movement Among the Classes

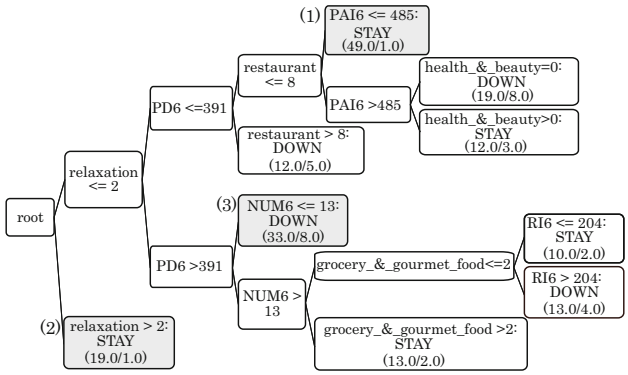

Fig. 6. Decision Tree which Classifies the Customers into STAY High Group and DOWN Group

could be a red flag that the customer will soon lose their interest in purchasing the coupon.

Similarly, Table 11 and Figure 7 illustrate the performance and the decision tree of LOW class model, respectively. From the decision tree in Figure 7, major rule antecedents that distinguish the UP class are emerged as shown grey node (1)∼(3). The explanatory variables such as gender, age, and RI6 are related to the rules. On the other hand, when male customers with lower PD (PD6 is large) and lower RI (RI6 is higher) satisfied the other purchasing conditions, they are predicted as the STAY class as shown in grey node (4). From these observations, we can deduce that this group of lower class customers are relatively young, curious, and tends to make impulsive purchases. In addition, this group of customers is not interested in purchasing homogeneous coupons repeatedly.

Based on the three cases mentioned above, in order to encourage customers to make repeated purchases for a prolonged period, it is important to entice

Table 11. Evaluation Index of the Decision Tree Model for the Lower Class

	Accuracy	Class	Precision	Recall	F-Measure	ROC Area
Training data	79.43%	UP	0.820	0.750	0.784	0.847
		STAY	0.773	0.838	0.804	0.847
Test data	59.93%	UP	0.606	0.550	0.577	0.612
		STAY	0.594	0.648	0.620	0.612

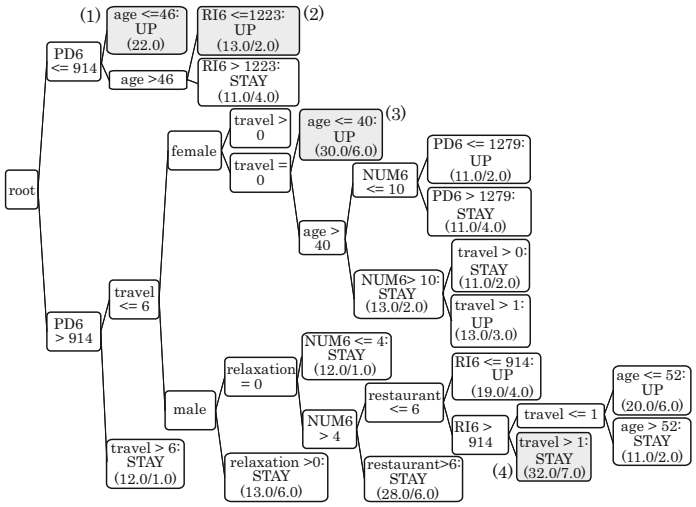

Fig. 7. Decision Tree which Classifies the Customers into STAY Lower Group and UP Group

them to purchase coupons from different categories. For higher class customers and customers moving up to a higher level, it is important to customize coupons tailored to their interests and curiosity drive. For example, in order to increase the number of active customers, the retailer should formulate a strategy to draw the customers' interest to purchase restaurant coupons for three or more times.

4 Conclusion and Future Works

In this paper, we have proposed three kinds of indices, popularity awareness index (PAI), purchase diversity (PD) and recency index (RI). By using these indices, customers can be segmented into several interesting classes, with corresponding analysis for each class. This approach is effective when the price of the product is irrelevant, which is applicable in this case. When considering other scenarios, this technique can be applied for analysis in other areas such as music CD market in which the price of the product is almost same.

As future works, the detail of movement of the 8 segments should be considered, which can be done by further dividing each index rank to reveal more

detailed attributes. In some cases, it may be effective to use only two indices of the three indices. We are also interested in applying our framework on the full set of real data in other areas, to substantiate the effectiveness of the model.

References

1. The Ministry of Economy, Trade and Industry in Japan: The FY2012 Research on Infrastructure Development in Japan's Information-based Economy Society (E-Commerce Market Survey) (2012)
2. Integrated service site for group buying coupon websites in Japan (COUPON-JP): `http://coupon-jp.com/`
3. Nomura, S., Oyama, S., Hayamizu, T., Ishida, T.: Analysis and Improvement of HITS Algorithm for Detecting WEB Communities. Journal of Systems and Computers 35(13) (2004)
4. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (1998)
5. Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM (JACM) 46(5) (1999)
6. Uno, H.: Ranking Techniques for web pages. Communications of the Operations Research Society of Japan 57(6) (2012) (in Japanese)
7. Gibson, D., Kleinberg, J., Raghavan, P.: Inferring Web Communities from Link Topology. In: Proc. of the 9th ACM Conference on Hypertext and Hypermedia (1998)
8. Rogers, E.M.: Diffusion of Innovations. Free Press (2003)