Original Article

Dietary life assistance for the university students by utilizing the meal prepaid card data

Toshihiko Shimauchi (Department of Regional Design and Development, Komatsu College, shimauchi@komatsu-c.ac.jp)
Daichi Fujita (Hokuriku Electric Power Company, jax24689@gmail.com)
Hidetaka Nambo (Graduate school of Natural Science & Technology, Kanazawa University, nambo@blitz.ec.t.kanazawa-u.ac.jp)
Haruhiko Kimura (Department of Regional Design and Development, Komatsu College, hkimura@komatsu-c.ac.jp)

Abstract

Recently, diversification of eating habits has been advancing in Japan. In such a situation, issues have been raised on the eating habits of younger generation, especially university students. The purpose of our study is to provide dietary life assistance to the university students by developing a system to locate students with unbalanced diet. Meal purchase patterns of university students are analyzed by utilizing purchasing history of the meal prepaid card users. The data are provided by a university cooperative association who runs university cafeterias. The number of the users analyzed are 2,458 (male 1,591 and female 867). The procedure of the analysis is as follows: k-means clustering is applied to the combination of food to generate 22 food clusters. Next PLSA is applied for clustering the customers to generate 8 customer clusters. For these two clusters, Bayesian network is applied to visualize correlation among factors. Based on this visualization and extraction of characteristics through clustering, customer clusters with eating habit problems are detected. Specific proposals for dietary life improvement are proposed to these clusters.

Key words

dietary life assistance, k-means, cross-validation, PLSA, Bayesian network

1. Introduction

Recently, diversification of eating habits has been advancing in Japan and issues have been raised on the eating habits of younger generation, especially university students. Many university students start to have more opportunities to eat out since they begin to live by oneself or by starting club activities or part time job which lead to wider socialization. As such, it is essential for university students to become independent from having his/her meal prepared by parents. However, according to a survey by the Cabinet Office of Japan, more than 40 % of students responded that they did not pay sufficient attention to their dietary life, and nearly 30 % of students answered they rarely had breakfast (CAO, 2009). Also, there are several studies pointing out that students do not take in sufficient nutrition. Meal skipping and unbalanced diet are regarded as one of the serious social problems, increasing the risk of life-style related diseases. To tackle these problems, many universities are taking various measures. For example, Kanazawa University has been holding various dietary education programs including soup sampling breakfast, cooking classes and dietary life counseling to raise awareness on healthy eating habits among the students (Kanazawa University Health Service Center, 2014). In addition, the university has been utilizing a meal prepaid card system. The card can be used in the university cafeterias. Students can deposit in advance certain amount of money into the card. By setting aside money for a meal, meal skipping can be prevented. The card system also enables to check a history of dietary life of the card holders, which makes it possible to advice healthier diet to those who have eating habit problems.

The purpose of our study is to provide dietary life assistance to the university students by developing a system to locate students with unbalanced diet. Usage history of meal prepaid card data provided by the university cooperative association are used to develop the system. The system employs several algorithm to cluster the data.

There are several existing studies concerning school/university cafeterias; surveys on cafeteria users (Goto et al., 2003b; Hitomi and Takaki, 2003); a study on student awareness on nutrition (Goto et al., 2003a); analysis of nutrient content (Kanda et al., 2012; Kamada et al., 2011); studies on menu selection (Ikeda et al., 2014; Namikawa et al., 2010; Shannon et al., 2002). Also there are several studies which propose balanced diet by clustering food selection of cafeteria users (Takechi, 2014; Nomura et al., 2015). Our study also proposes a method for dietary life assistance by clustering cafeteria users. However, Takechi (2014) does not use ID-POS, so it is not possible to obtain time series data for each specific user, precluding the possibility to assist each student according to his/her dietary life. For the difference from Nomura, Nambo and Kimura (2015), our study employs PLSA as a clustering method, which allows to analyze co-occurrence matrix. Hence more accurate clusters are obtained. After obtaining the clusters, Bayesian network is applied to locate customers with unbalanced diet to whom specific assistance are proposed.

2. Proposed method

This chapter explains the procedure of the method proposed in our study. Figure 1 shows the flowchart of the procedure.

Figure 1: The flow of proposed method

2.1 Outline

This section outlines the method of our study.

First, from the data provided, information on purchased items for each meal are extracted. They are converted to food categories based on the middle classification code in the data to generate a dataset for our analysis.

Second, the combinations of purchased food categories in this dataset are used as attributes for k-means clustering to generate clusters of "purchasing pattern for food categories per one meal (hereafter referred to as food cluster) which shares similar purchasing patterns. After this clustering, percentage of food clusters purchased by each customer are calculated to generate another dataset consisting of each customer as rows and purchasing percentage of each food cluster as columns.

This new dataset are used to apply PLSA and cluster customers (hereafter referred to as customer cluster).

Finally, Bayesian network is generated by using the result of the customer clusters and the food clusters mentioned above. By analyzing the network, customers with unbalanced diet are located and specific assistance are proposed.

Following sections make detailed explanations on each step outlined above.

2.2 Generating datasets

This section explains the data used in our study and the procedure to generate datasets.

2.2.1 Data used

Our study uses purchasing history data stored in the meal prepaid cards. The data are provided by Kanazawa University Cooperative Association. The customers studied are 2,458 students who have usage history out of 2,474 students from first to third grade with meal prepaid card contract. Detailed breakdowns are shown in Table 1.

Table 1: Students surveyed

Sex	Male 1,591	Female 867	
Grade	G1 818	G2 811	G3 819

The period of study is 30 months starting from April 2014 to September 2016. The total number of the record in the data used are 1,416,581. The data is composed of 16 attributes: user code, sex, code of cafeteria used, date of purchase, time of purchase, middle classification code, item code, item name, the number of items purchased, the amount of money, calories, three color codes(red, green and yellow, which will be explained later), and salt intake. The multiple records with the same user code, the date of purchase and the time of purchase signify that these purchases are done at the same time; hence these records are aggregated as the record per one meal.

2.2.2 Converting food category

As explained in previous section, the date provided include 2,458 customers and 461 food items. Generating a dataset without any pretreatment from these data will lead to a dataset with sparse matrix and makes it difficult to find any patterns nor similarities by analyzing it. To overcome this problem, the food items are converted into 17 food categories by using the middle classification codes included in the original data. 17 food categories are as follows: rice (low), rice (high), bowl (low), bowl (high), main dish (low), main dish (high), side dish (low), side dish (high), noodle (low), noodle (high) other noodles, dessert, beverage, self-servicing bar, rice and noodle, set menu and takeout lunch. For 5 items (rice, bowl, main dish, side dish, and noodle), which have larger number of items in the food categories and have wide variety of price range, each item is divided into two sub-items according to the unit price (high and low). The threshold of the division is a mean score of each item. The item with higher than mean price is designated as high and the item with lower price as low.

2.3 Food clustering

This section explains the purpose and the process of food clustering.

One of the objectives of our study is to understand the current status of dietary life of university students. It is possible to calculate the purchase probability of each food category by every customer by using the data described in the preceding section and to cluster the customers. This method will clarify the rate of food category purchased for each customer cluster. However, it is not possible to understand how each customer chooses a combination of food categories per one meal; hence the understanding of the dietary life is difficult to achieve. In order to solve this problem, before clustering the customers, the selection and combination of food per one meal are clustered.

K-means algorithm is applied for clustering. Before applying k-means to any given dataset, the value of k should be determined. In order to determine the value, ABC analysis is conducted to all the 387 patterns of food category combination. ABC analysis is a method to divide the data into three categories A, B and C with A being most important. Data are arranged according to the importance for given analysis. For our analysis, the importance is defined as a pattern of food combination for one meal. The result of the analysis leads to the 22 patterns of food combination included in group A. So k value of 22 is adopted.

Based on the result of the food clustering, the rate of selection for food cluster by each customer is calculated. The result of the calculation is used as a dataset for the customer clustering in the following section.

2.4 Customer clustering

Based on the dataset described in the preceding section, PLSA (Probabilistic Latent Semantic Analysis) is applied to cluster the customers. The number of the clusters is determined by normalized mean absolute error which will be described later. PLSA is a soft clustering method and data can belong to multiple clusters. However, one of the objectives of our study is to assist dietary life of each customer. To achieve this objective, a cluster with the highest membership probability is designated as the customer cluster for our analysis.

2.4.1 Evaluation method

The number of the cluster must be determined in advance for the customer clustering. Normalized mean absolute error of the number of the items purchased for each food category by every customer cluster is used as evaluation index and leave-one-out cross validation is used for the evaluation method in order to determine the appropriate number of clustering. For the evaluation index, F-measure or perplexity are often used. However, in order to calculate F-means, correct labels are required for the test data. Our study does not have "correct combination" for the customer and the food, so this method is not appropriate for our study. For the perplexity, it is originally designed for text mining, so the method is not appropriate for our study. Hence normalized mean absolute error for the number of items purchased for each food category by each customer cluster is used.

Next, the method to calculate normalized mean absolute error is explained. For the customer who belongs to customer cluster k, the number of the items purchased which belongs to food category l is N_{kl}. Also, for the training data, the number of the items purchased is $N_{kl}^{train,}$ and the number of the items purchased in test data is N_{kl}^{test}. In our study, N_{kl}^{train} is the total number of purchase for 29 months, and N_{kl}^{test} is the number of items purchased for one month. Differences in duration and the size of data make it impossible to directly compare residuals between the training data and the test data. For the purpose of normalization, the training data are divided by the number of the all purchases for 29 months and the test data are divided by the number of the purchases for remaining one month. The fitness of the clustering is calculated by using normalized mean absolute error E_{kl} between N_{kl}^{train} and N_{kl}^{test}. E_{kl} is defined as (1).

U stands for the number of customer clusters and V stands for the number of food categories. Hence U = 2; 3; :::; 30 and V = 17.

$$E_{kl} = \frac{N_{kl}^{train}}{\sum_{k=1}^{U} \sum_{t=1}^{V} N_{kl}^{train}} - \frac{N_{kl}^{test}}{\sum_{k=1}^{U} \sum_{t=1}^{V} N_{kl}^{test}} \qquad (1)$$

Normalized mean absolute error E is defined by calculating the mean value for all the customer clusters and food categories. And this value is used for evaluating the clustering methods used in our study. The formula to calculate normalized mean absolute error is as follows:

$$E = \frac{1}{UV} \sum_{k=1}^{U} \sum_{l=1}^{V} E_{kl}$$

The result of the error derived from the above mentioned calculation is evaluated to calculate the similarity between N_{kl}^{train} and N_{kl}^{test}.

2.4.2 Evaluation results and determination of the number of clusters

Figure 2 shows the graph of normalized mean absolute error with the number of customer clusters ranging from 2 to 30. As for the value of Tempered EM, 0.90, 0.95 and 0.99 are used. For each cluster, the value 0.90 leads to the minimum error. Hence the value of 0.90 is adopted.

Figure 2 shows as the number of the clusters increases, the value of the errors decreases. But the rate of the decreases in the value of the errors shrink as the number of clusters increases. So when the number of the cluster increases, the errors will decreases accordingly, but the model based on the analysis becomes more complex and makes it difficult to understand the characteristics of each cluster. Also, T-analysis is applied for the cluster number 2, 8, and 30. There are significant difference between 2 and 8, but not between 8 and 30. Adding to these two points, the existing study by Nomura et al. (2015) employs 8 as the number of the customer clusters. Our study also employs 8 so that the results can be easily compared with the existing study. The normalized mean absolute error for the existing study is 0.2550 % and for our proposal is 0.1525 %. It is clearly shown that our proposed method has improved the accuracy.

3. Organizing the cluster characteristics

Employing the methods explained in chapter 2 generates customer clusters U_1-U_8 and food clusters S_1-S_{22}.

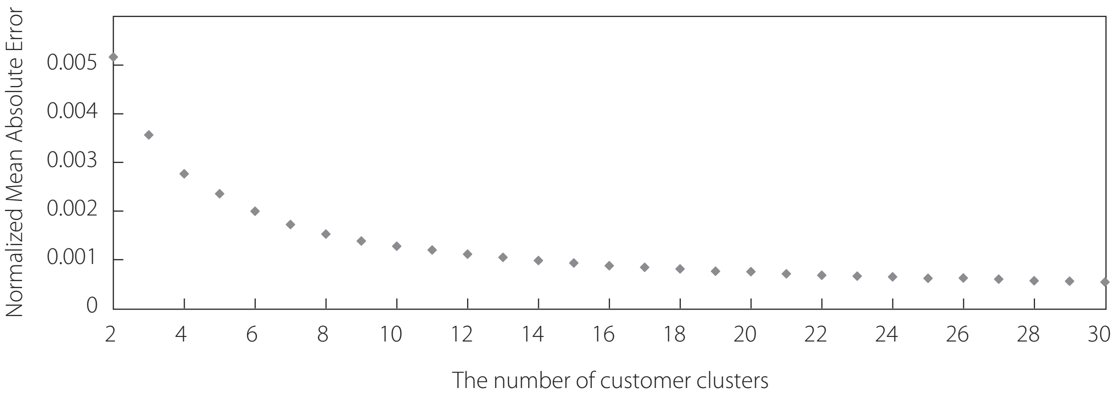

Figure 2: Normalized mean absolute error

3.1 The characteristics of customer cluster

In this section, the characteristics of each customer cluster are organized. Table 2 shows the number of the customers and the male/female ratio for 8 customer clusters generated. The table shows that clusters U_2, U_4, U_5, and U_8 are male dominant, and U_3 and U_7 are, to the contrary, female dominant. For the clusters U_1 and U_6, there are small differences in male and female students. From these observations, it can be said that the result of the clustering better reflects the food selection by male than by female.

Table 3 shows the average number of items purchased, the average amount of money spent, and average calories per one meal for each customer cluster dating from April 2014 to September 2016. The table shows U_4 and U_8 are the clusters with higher average amount of money spent and the average calories. To the contrary, U_3 and U_6 are the clusters with less average amount of money spent and the average calories.

In terms of the average number of items purchased, U_2, U_3 and U_6 are the clusters with small value. A dietary guideline published by the Ministry of Agriculture, Forestry and Fisheries (MAFF, 2017) advises for better food combinations which are centered on staple food, main dish and side dishes. Based on this guideline, these 3 clusters have several rooms for improvement for dietary life.

Table 2: Average of feature value of customer clusters (1)

Customer cluster	Number of users	Male	Female	Grade 3	Grade 2	Grade 1
U_1	464	224	240	143	159	162
U_2	232	203	29	70	71	91
U_3	530	125	405	198	155	177
U_4	339	335	4	111	135	93
U_5	312	303	9	77	84	151
U_6	74	50	24	22	21	31
U_7	259	112	147	68	108	83
U_8	248	239	9	129	78	41

Table 3: Average of feature value of customer clusters (2)

Customer cluster	Purchased dishes	Money (yen)	Calories (kcal)
U_1	3.66	413.34	698.65
U_2	3.35	431.14	754.21
U_3	2.93	354.54	580.19
U_4	4.31	474.98	881.67
U_5	3.67	434.53	822.22
U_6	3.20	394.60	658.17
U_7	3.64	403.08	677.20
U_8	3.72	455.13	809.82

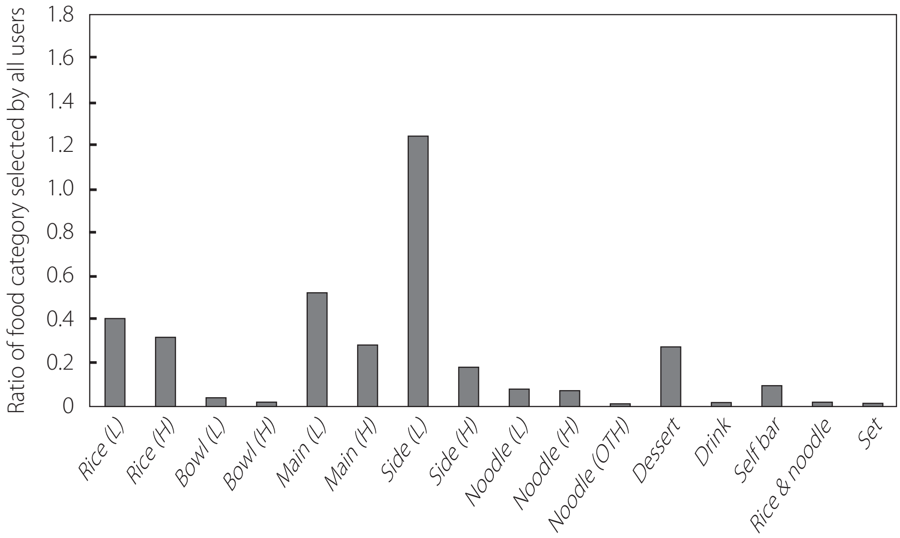

Figure 3: Ratio of food cluster selected by all users

Figure 3 shows the average number of items purchased for each food category by all customers. Table 4 shows the characteristics of each customer cluster.

3.2 Characteristics of food cluster

In this section, the characteristics of each food cluster generated are organized. Table 5 shows the number of data included in food clusters from S_1 to S_{22}.

Cluster S_4, S_{11} and S_{17} include larger number of data. The purchasing pattern included in these clusters are popular among the customers surveyed. To the contrary, clusters S_2, S_{10}, S_{14}, and S_{21} have smaller number of data, which shows these patterns are less popular among the students.

Next, the nutritious balance for each food cluster is analyzed. Based on the dietary guideline published by the MAFF, a purchasing pattern with 4 items or more can be evaluated as balanced diet.

Apart from this guideline, there is another standard for measuring nutrition intake. The standard is called three food-group scoring (NFUCA, n.d.). In this scoring, daily necessary intake are divided into three food groups: red (protein), green (vegetables) and yellow (carbohydrate and fat). For each food group, 80 kcal is designated one point. For example, 350 grams of vegetables are translated into one point of green. Without detailed nutrition knowledge, one can maintain a balanced eating habit by following this scoring system. The recommended daily scores for a male student are 3 points of green, 6 points of red and 16 points of yellow, which can be translated into 1 point of green, 2 points of red and 5 or more points of yellow for a meal.

The cafeterias surveyed for our study employ this scoring system and show these information on the menu card for every meal. Customers can choose their menu by referring to these nutrition information.

Based on this guideline, food clusters listed in Table 5 are assessed. In general, points of green are not sufficient since no food clusters reach recommended intake of one point. Green food-group consists mainly of vegetables. This observation coincides with an official annual survey showing younger generation do not take in sufficient vegetables. (Ministry of Health, Labour and Welfare, 2017). For the food clusters S_2, S_5, S_8, S_{10}, S_{11} and S_{19}, they mostly satisfy three color

Table 4: Property of each customer cluster

Customer cluster	Property
U_1	Order rice (l) and many main/side dishes
U_2	Order various items without any characteristics, but with less items purchased
U_3	Female dominant cluster with less money spent and calorie taken
U_4	Cluster with more money spent and calorie taken
U_5	Order rice (h) and two side dishes
U_6	Tend to prefer noodles and purchase less items
U_7	Order rice(l).Sometimes order dessert
U_8	Cluster with average dietary lifestyle

Table 5: Average of feature value of each food cluster

Food cluster	Number of data	Money (yen)	Calories (kcal)	Red (pt)	Green (pt)	Yellow (pt)	Salt (g)	Purchased dish
S_1	7636	489.33	860.21	1.66	0.44	8.69	3.13	4.05
S_2	2199	536.46	742.74	2.14	0.65	6.51	3.81	4.99
S_3	7330	476.85	641.10	1.82	0.53	5.69	3.24	4.43
S_4	36526	332.95	660.24	1.88	0.58	5.81	3.41	3.50
S_5	19635	538.35	1116.14	2.61	0.71	10.66	4.61	4.17
S_6	13059	289.77	520.37	0.86	0.34	5.32	4.89	2.02
S_7	5327	347.35	404.49	1.01	0.26	3.80	2.94	2.43
S_8	19523	463.82	809.90	2.10	0.75	7.30	3.86	4.73
S_9	4869	407.65.	711.40	1.83	0.34	6.72	3.25	2.03
S_{10}	1351	591.05	924.04	2.44	0.70	8.42	4.65	3.76
S_{11}	25880	453.28	1005.87	2.15	0.71	9.75	4.22	4.15
S_{12}	5581	525.90	1097.06	2.73	0.72	10.33	4.82	4.56
S_{13}	6334	606.96	1187.39	2.41	0.84	11.63	4.54	5.43
S_{14}	2497	488.94	1016.23	3.62	0.80	8.30	5.62	4.49
S_{15}	14356	373.35	674.06	1.15	0.36	6.92	6.83	1.83
S_{16}	12189	422.93	978.38	2.03	0.54	9.70	3.64	3.88
S_{17}	33753	431.34	778.49	2.34	0.65	6.76	4.01	3.93
S_{18}	19233	264.89	498.70	1.72	0.48	4.04	2.95	2.43
S_{19}	15606	351.09	678.56	2.00	0.62	5.91	3.67	4.18
S_{20}	5282	253.75	368.69.	0.74	0.19	3.69	1.44	2.72
S_{21}	3269	582.19	894.80	2.04	0.61	8.57	4.42	4.58
S_{22}	7632	399.39	726.18	1.58	0.47	7.04	3.36	2.67

points with sufficient number of items purchased. They are food clusters with balanced nutrition. For S_1 and S_3, green points are not sufficient but the customers in these clusters use self-servicing bar. The bar does not show the food point since various food groups are offered. By converting the money spent on the bar into the amount of vegetable intake, S_1 and S_3 can be judged to have balanced nutrition. Clusters S_6, S_7, S_{15}, S_{18} and S_{22} are low in the number of items purchased and can be judged as lacking enough nutrition intake.

For the amount of salt intake, the ceiling of the daily intake is set at 8 grams. Clusters S_5, S_6, S_{10}-S_{15} and S_{21} exceed the 4 grams of salt intake per one meal and can be judged to have a risk of the excessive intake.

As a summary for this section, Table 6 overviews properties of each food clusters and Table 7 shows the evaluations for nutrition balance of each food cluster.

3.3 Visualization of factor relationships by using Bayesian network

In this section, Bayesian network is employed to visualize the relationship among factors in food clusters, customer clusters and customer variables. Our study uses 1,416,581 effective records and 2,458 effective customers. These records and the customers are designated the information about customer clusters, food clusters, food categories and customer variables. Customer variables are sex, years, average amount of money, frequency of the cafeteria usage, average number of items purchased and average calories. Sex and years are dummy variables and designated 1 when the data is present and 0 when the data is absent.

The result of the visualization of the factors is shown in Figure 4. Figure 5 is derived from Figure 4 and shows a partial network focusing on students' grade.

From Figure 5, the characteristics for each grade are summarized as follows: The first grade customers tend to choose clusters S_1, S_5, S_7 and S_{10}. For the third grade customers, S_3, S_9, S_{20}, and S_{21} are popular.

The first grade students can be divided into two groups: one with good dietary life with balanced nutrition by choosing food clusters S_5 or S_{10} and the other with unbalanced dietary life by choosing food cluster such as S_7 which signifies less balanced nutrition.

On the other hand, for the third grade students, all of them choose food clusters with certain amount of risk for unbalanced nutrition.

Table 6: Property of food clusters

Food cluster	Property
S_1	Good nutrition balance with rice(h) and sufficient vegetables by main or side dish
S_2	Good nutrition balance with rice(l) and sufficient side dishes
S_3	Good nutrition balance by using self-servicing bar
S_4	Order rice, one main and one side.
S_5	Order rice and sufficient main and side. The highest calorie intake but good nutrition balance
S_6	Often order noodle only. Less purchased items. Concern for nutrition balance.
S_7	Although using self-servicing bar, with less purchased items. Concerns for nutrition balance.
S_8	Good nutrition balance with many items purchased
S_9	Order rice bowl(h) and one side dish. Less purchased items and concerns for nutrition balance.
S_{10}	Order rice bowl(h) and two side dishes. Good nutrition balance but high in salt intake.
S_{11}	Good nutrition balance with rice(h) and many vegetables by main/side dish
S_{12}	Order rice(h) and take in many vegetables by main/side dish but very high in salt intake
S_{13}	Order rice(h) and take in many vegetables by main/side dish but very high in calorie intake
S_{14}	Order rice(l) and take in many vegetables by main/side dish but very high in salt intake
S_{15}	Order rice(h) often with less purchased items. Bad nutrition balance
S_{16}	Order rice(h) and two dishes from main/side dish. Low in green score.
S_{17}	The largest cluster with good balance for rice and side dish. High in salt intake
S_{18}	Typical cluster for those who do not order rice
S_{19}	Good nutrition balance with rice(l) and three items from main/side. Many items purchased
S_{20}	Bad nutrition balance with low calorie intake with rice(l) and self-servicing bar
S_{21}	Order various items but high in salt intake
S_{22}	Bad nutrition balance with bowl and side dish. Less items purchased.

Table 7: Nutrition balance of each food cluster

Nutrition balance	Food clusters
Good	S_2, S_5, S_8, S_{10}, S_{11}, S_{19}
Partly good	S_1, S_3, S_4, S_9, S_{12}, S_{13}, S_{14}, S_{16}, S_{17}, S_{20}, S_{21}
Bad	S_6, S_7, S_{15}, S_{18}, S_{22}

4. Dietary life assistance proposal

In this chapter, several proposals on dietary life are proposed based on the result of characteristic organization of food clusters explained in chapter 3.

4.1 Assistance to the customer cluster U2

Figure 6 shows a partial graph for the customer cluster U2.

Figure 6 shows that from U_2 runs directed edges to S_5 and S_6, meaning that a customer in U_2 has high probability to choose purchasing patterns of S_5 or S_6.

Food cluster S_5 is characterized as sufficient number of items purchased (4.17) and good scores on red (2.61) and yellow (10.66), but a low score in green (0.71) and high calorie intake(1116.14 kcal).

Food cluster S_6 is characterized as insufficient number of items purchased (2.02) and low scores on red (0.86) and green (0.34). Calorie intake is low (520.37 kcal).

From these observations, dietary problems of a customer in U_2 can be summarized as follows:

- In terms of the dietary guideline which advised to have many items for each meal, a customer in cluster U_2 satisfies the guideline when they choose S_5. However, when S_6 is chosen, the meal is mainly centered on noodle which leads to an unbalanced diet.
- In terms of three food-group score method, a customer cluster U_2 satisfies yellow score, but low in green score.

Based on this summary, food assistance to a customer in cluster U_2 is proposed. A customer in this cluster does not take in sufficient vegetables. Also, when he/she chooses S_6 and have noodle centered meal, his/her diet becomes unbalanced. The customer is advised to take part in a vegetable soup sampling run by the university to make up for low green score. Also, the customer is advised to add one more side dish for his/her meal when choosing noodle centered meal.

4.2 Assistance to customer cluster U4

Figure 7 shows a partial graph for the customer cluster U4.

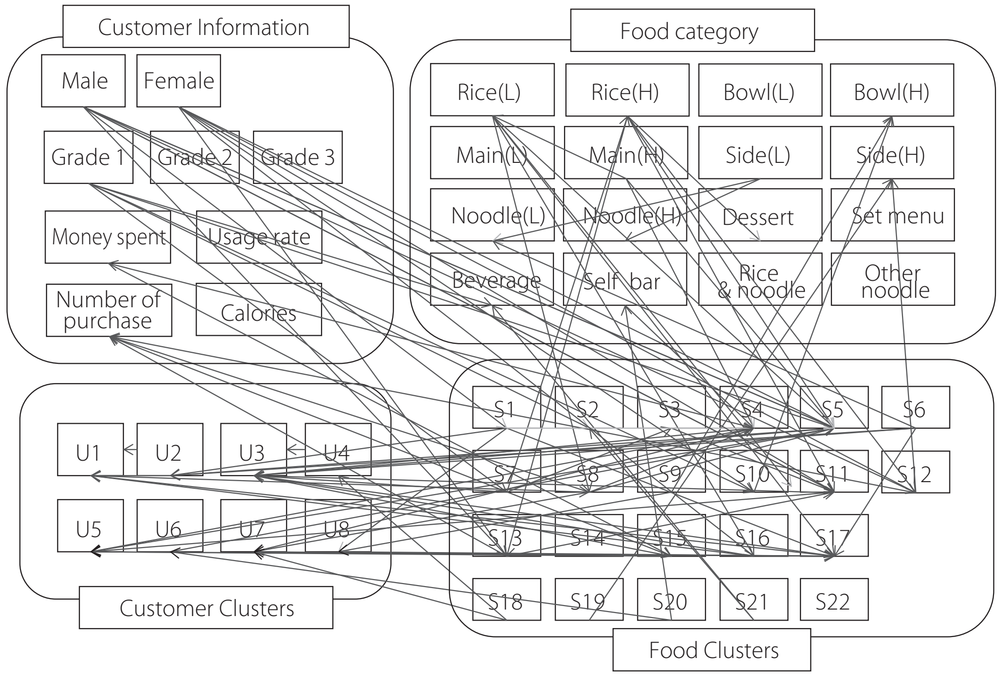

Figure 4: Bayesian network

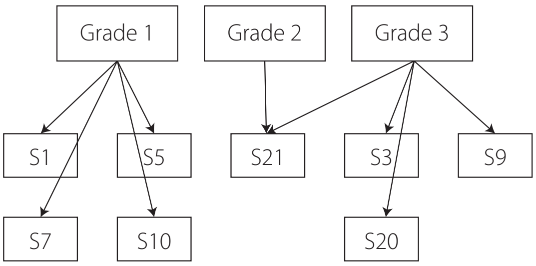

Figure 5: Partial network focused on students' grade

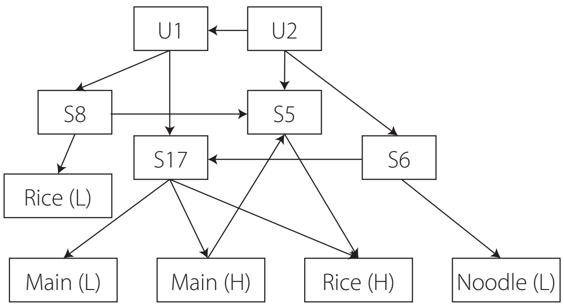

Figure 6: Partial network of customer cluster U2

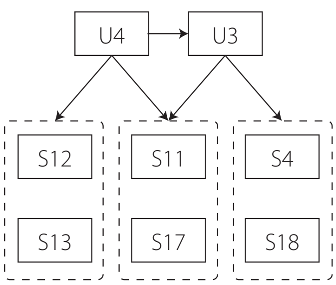

Figure 7: Partial network of customer cluster U4

Figure 7 shows that from U_4 runs directed edges to S_{12}, S_{13}, S_{11} and S_{17}, meaning that a customer in U_4 has high probability to choose purchasing patterns of these four clusters.

Food clusters S_{11}, S_{12} and S_{13} share similar characteristics. They are all sufficient in the number of items purchased (4.15, 4.56, 5.43) with good scores on red (2.15, 2.73, 2.41) and yellow (9.75, 10.33, 11.63). They are all low in green (0.71, 0.72, 0.84) and high in calorie intake (1005.87 kcal, 1097.06 kcal, 1187.39 kcal).

Food cluster S_{17} is characterized as sufficient number of items purchased (3.93) and good scores on red (2.34) and yellow (6.76), but a low score in green (0.65). The calorie intake is moderate (778.49 kcal).

From these observations, dietary problems of a customer in U_4 can be summarized as follows:

- In terms of the dietary guideline, a customer in cluster U_4 satisfies the guideline for any choice of food cluster.
- In terms of three food-group score method, a customer cluster U_2 satisfies yellow score, but low in green score.
- In terms of calorie intake, the customer tend to take in excessive calorie except for S_{17}.

Based on this summary, food assistance to a customer in cluster U4 is proposed. A customer in this cluster tends to choose high calorie meal. In order for the customer to be aware of his/her calorie intake, it is advisable to modify the meal card system so that the user can view recommended meal from "my page" which can be accessed on the internet and shows usage history for each month. Also, the customer is advised to choose the food cluster S_{17} with lower calorie intake menu.

5. Conclusion

Our study proposes dietary life assistance by developing a system to locate students with unbalanced diet. Two staged clustering is applied: one for meal purchasing patterns of university students to generate food clusters and the other for customers with similar food cluster preference to generate customer clusters. This two staged clustering aims to improve accuracy of the prediction on the purchasing pattern of each customer cluster. Using the proposed method, customer clusters U_1 to U_8 and food clusters S_1 to S_{22} are generated. Analyzing the clusters clarify the selection rate of food categories and nutrition intake for each customer cluster. Specifically, for every food cluster, the score for green is low regardless of the number of items purchased. Also, according to the grade, the dietary life shows difference: newly enrolled students tend to choose balanced diet, but third grade students tend to choose unbalanced diet. Based on the findings through Bayesian network, dietary life assistance to customer clusters U_2 and U_4 are proposed.

Our study analyses the data totaling 1,416,581 collected for 30 months period. They are sufficient to develop a robust system that can locate students with unbalanced diet and propose specific advice. However, the scope of the analysis is limited to one specific university. When the proposed method is applied to another dataset collected at different university, the system may generate clusters which differ from our study. Expanding the scope of the study by collaborating with scholars from different universities can generate a comprehensive system.

Acknowledgement

We would like to extend our deepest appreciation to Mr. Kita Toshiyuki, a managing director of the Kanazawa University Cooperative Association, Ms. Inoue Misako, an assistant to the managing director and other staff of the cafeterias for their invaluable support to our study.

We are scheduled to submit an application for research involving human subjects to the medical ethics committee of Kanazawa University where the students in this study enrolled.

References

Cabinet Office, government of Japan (2009). Research report on state and opinion of university students for dietary life. 1-76.

Goto, Y., Fujimoto M., and Hamana, T. (2003a). A survey on students' dietary habits at the university cafeteria. *Bulletin of the Faculty of Education, Yamaguchi University*, Vol. 52, No. 1, 35-50.

Goto, Y., Oishi, N., Takenaka, R., and Furukawa, K. (2003b). A survey on university students' breakfast: Focusing on their dietary behavior. *Bulletin of the Faculty of Education, Yamaguchi University*, Vol. 53, No. 1, A31-A50.

Hitomi, E. and Takaki, M. (2009). Investigation of Yamaguchi Prefectural University cafeteria and an action for the improvement. *Archives of Yamaguchi Prefectural University*, Vol. 2, 53-61.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57.

Honda, A., Nakamura, O., and Kai, Y. (2010). Study on college students' dietary life and attitude toward prevention of lifestyle-related diseases. *Journal of Environmental Studies, Nagasaki University*, Vol. 12, No. 2, 89-96.

Ikeda, M., Onezawa, H., Kamisaka, N., Takahashi, K., Mochizuki, N., Hirasawa, M., Seki, C., and Sawayama, S. (2014). Gender difference of university students depends on their tendency to choose dishes in cafeterias. *Journal of Japanese Society of Shokuiku*, Vol. 8, No. 1, 9-17.

Kamada, C., Yoshimura, Y., Okumura, R., and Takahashi, K. (2011). Examination of the validity of SV calculation based on "Japanese food guide spinning top" in university student. *Journal of the Japan Dietetic Association*, Vol. 54, No. 1, 17-27.

Kanazawa University Health Service Center (2014). Student support program. *Bulletin of Kanazawa University Health Service Center*, Vol. 40, 33.

Kanda, T., Matsuzaki, A., Tokumoto, M., Yoshikawa, M., Yuzuriha, C., Kubota, M., and Okumura, H. (2012). Analysis of nutrient- and energy-intakes in lunches taken by male university students at a university cooperative cafeteria: Comparison of users and non-users of meal cards. *Journal of Japan School Meal Management*, Vol. 6, No. 2, 51-64.

Machine Learning Group at the University of Waikato. "Weka 3-data mining with open source machine learning software in Java" retrieved from http://www.cs.waikato.ac.nz/ml/weka/ (accessed on 2016/01/22).

Ministry of Agriculture, Forestry and Fisheries (2017). Dietary guideline. http://www.maff.go.jp/j/syokuiku/shishinn.html (accessed on 2017/7/31).

Ministry of Health, Labour and Welfare (2017). National health and nutrition survey.

Namikawa, S., Taniwaki, A., and Yamakita, H. (2010). A consideration of action concerning university students' eating habits, and an investigation concerning menu selection etc. in school canteens. *Journal of Human Developmental Research*, No. 1, 39-48.

National Federation of University Co-operative Associations (n.d.). What is three food-group scoring? http://www.univcoop.or.jp/service/food/three.html (accessed on

2017/7/31).
Nomura, K., Nambo, H., and Kimura, H. (2015). Classification of school cafeteria users using the meal prepaid card data. *IEEJ Transactions of Electronics, Information and Systems*, Vol. 135, No. 10, 1262-1269.
Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings of the 7th Conference of the Cognitive Science Society*, 329-334.
Raschka, S. (2016). *Python machine learning*. Impress.
Shannon, C., Story, M., Fulkerson, J., and French, S. (2002). Factors in the school cafeteria influencing food choices by high school students. *Journal of School Health*, Vol. 72, No. 6, 229-234.
Shinnou, H. (2007). *Learning cluster analysis*. Ohmsha.
Takechi, S. (2014). Approach of healthcare administration utilizing purchase data of school cafeteria. *Proceedings of the Asia Pacific Industrial Engineering & Management Systems Conference 2014*, 594-601.

(Received: June 21, 2017; Accepted: October 16, 2017)