Valuation of Customer and Purchase Behavior of a Supermarket Chain Using ID-POS and Store Causal Data

Syun Usami[1(✉)], Kohei Otake[2], and Takashi Namatame[2]

[1] Graduate School of Science and Engineering, Chuo University,
Hachioji, Japan
al3.nnma@g.chuo-u.ac.jp
[2] Faculty of Science and Engineering, Chuo University, Hachioji, Japan
{otake,nama}@indsys.chuo-u.ac.jp

Abstract. Along with the growth of internet market a shopping, retailers such as supermarket chains need a strategy corresponding to customers in each store. The purpose of this research using ID-POS data of supermarket chain is to clarify customer characteristics and purchase behavior for each store. First of all, we categorize stores based on causal data concerning each store such as sales floor area and peripheral population. Second, we analyze customer's purchasing behavior using ID-POS data for each class that we tried to classify above, and extract characteristic purchasing behavior. Finely we evaluate these results together and clarify customer characteristics and purchasing behavior for each store causal.

Keywords: POS · Store causal data · Association analysis · Decile analysis · Hierarchical cluster analysis

1 Introduction

Recently, EC (Electronic Commerce) sites that are online stores to purchase variety of products through internet getting popular (Fig. 1). Under such situation, in real-world retailing such as supermarket industry, are required improvement of sales by unique service.

In order to propose some new unique services, analyst must know the characteristics of customers and tendency of purchasing behavior. Moreover, characteristics of each store are equally important to consider because the characteristics of store such as floor layout, competitor of the area and population of the market area, are very different. However, previous analyses of real-world retailing performed only purchasing history. It is necessary analyze using store causal data to improve the sales because customers are different depending on the surrounding environment.

Fig. 1. Global online retail sales have increased 17% yearly since 2007 [1]

2   Purpose of This Study

In this study, we aim to clarify customer characteristics and purchasing behavior using store causal data. It can provide useful suggestion for proposing sales strategies such as shelf allocation within stores and types of products to be stocked compared to analysis using only sales history. In addition, analysis of the relationship between product sales histories, will lead to improvement in store sales and satisfaction.

3   Analytical Procedure

We use ID-POS (point of sales with customer ID number) data of 53 store at a supermarket chains from March 2015 to February 2016 excluding nonmember data. Figures 2 and 3 the summarizing own data.
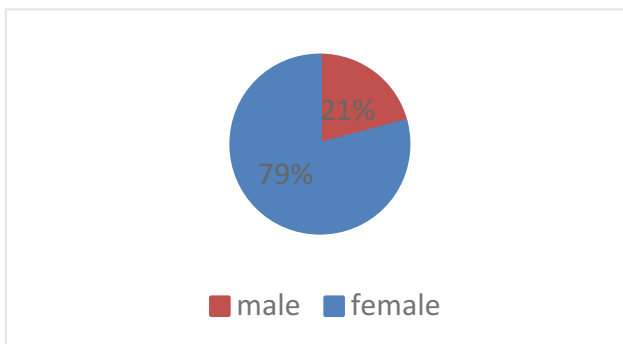

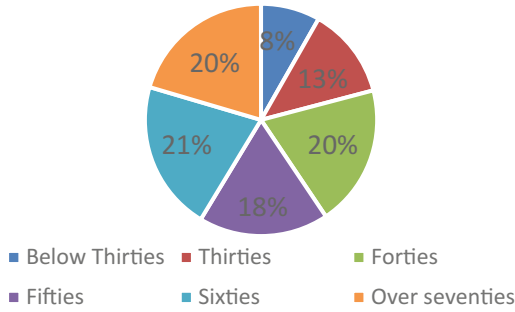**Fig. 2.** Male-female ratio of all customers

Fig. 3. Compared ratio of age

Figure 4 shows the outline of analytical procedure. First, we classify stores by cluster analysis using store causal data. Second, focusing on each classified cluster, we aggregate customer characteristics such as age structure and household composition and classify customers. Second, we perform association analysis on purchasing behaviors and clarify concurrent selling commodities for each cluster. Finally, we clarify customer characteristics and purchasing behavior for each store causal from these analysis results.
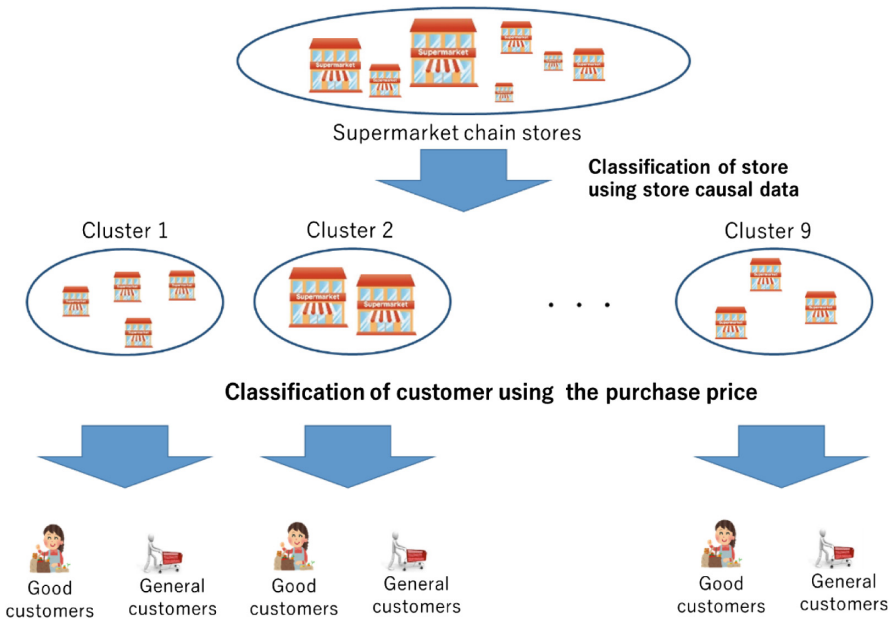

Fig. 4. Outline of analytical procedure

4   Analysis Method

4.1   Hierarchical Clustering Analysis

Hierarchical cluster analysis is a method of classifying objects by creating collections of objects similar to each other from a group in which objects of different properties are mixed based on the distance between objects.

In this study, 53 stores were classified into nine clusters by using store causal data, in order to clarify the relationship between the characteristics owned by the store and the customer characteristic and purchasing behavior.

The cosine distance was used as the distance between objects used for classification, and the Ward method [2] was adopted to merge cluster (Table 1).

Table 1.   Variables used for hierarchical clustering

Variable name	Summary
The maximum parking number	The maximum number of parking lots of each stores
Sales floor space	Sales floor space of each stores
City center	Based on the address of the store, three dummy variables were created from the surrounding information on the map
Suburbs	
Mountainous areas	Apply: 1 Not apply: 0
The number of items	POS data with ID of each stores
The population in commerce areas	A circle with the diameter of the cube root of the maximum parking number centering on the store was set as a set trade area, and the population of the market area was totalized

4.2   Decile Analysis

Decile analysis is an analytical method for calculating the sales composition ratio of each rank by ranking the purchase price of all customers, based on purchase history data.

Generally, decile analysis is conducted to clarify customers with high purchase price per cluster and to evaluate differences in purchase behavior with other customers.

In this study, we performed decile analysis for each cluster that classified by hierarchical cluster analysis, and we defined that Decile rank 1 to 8 customers as "general customers" and 9 to 10 customers as "good customers".

4.3   Association Analysis

Association analysis or market basket analysis in marketing is used to extract meaningful relevance between products from enormous log data. In order to pick up concurrent selling relationship and to obtain suggestions leading to sales measures such as display and sales floor placement, association analysis was conducted.

In this analysis, association analysis is performed for "good customer" and "general customer" set for each cluster, we used the basket ID assigned to the shopping cart as a key for making concurrent selling relationship.

5 Results of Analysis

5.1 Store Classification

The results of the hierarchical cluster analysis are shown in Tables 2 and 3.

Table 2. Result of hierarchical cluster analysis (1)

Cluster	Number of stores	Average parking number	Average sales floor space	Average number of items	Average population in commerce areas
1	6	49.5	852.5	21794.7	56798.7
2	5	76.2	958.1	23726.8	69637.0
3	9	68.3	734.7	19412.3	33160.4
4	3	733.3	1524.6	28499.7	11721.0
5	5	131.8	1137.3	25644.8	14143.8
6	10	148.6	1346.5	27322.7	51253.8
7	6	170.0	1377.6	26993.8	107992.8
8	5	141.2	1569.1	29773.2	123128.0
9	4	387.8	854.5	17575.3	176676.0

Table 3. Result of hierarchical cluster analysis (2)

Cluster	Percentage of city center stores	Percentage of suburban stores	Percentage of mountainous areas stores
1	100%	0%	0%
2	0%	100%	0%
3	0%	100%	0%
4	0%	0%	100%
5	0%	0%	100%
6	0%	100%	0%
7	0%	100%	0%
8	100%	0%	0%
9	50%	50%	0%

- Clusters 1 and 8
  They composed only of city center stores.
- Clusters 2, 3, 6 and 7
  They composed only of suburban stores, clusters 4 and 5 are clusters constituted only of mountainous stores.
- Cluster 9
  It composed of city center stores and suburban stores.

In order to extract customer features for each cluster, we compiled the ratio of male and female (Fig. 5), the ratio of age, the ratio of unmarried and married (Fig. 6), and the ratio of the number of household members.
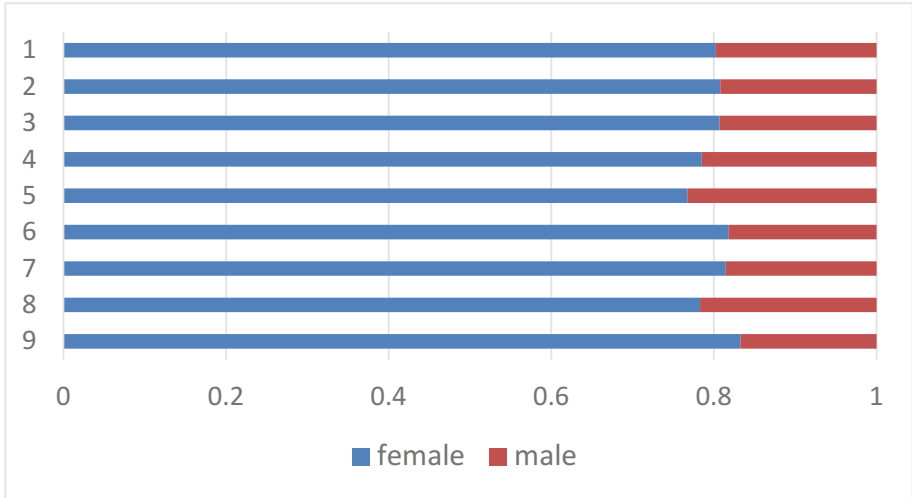

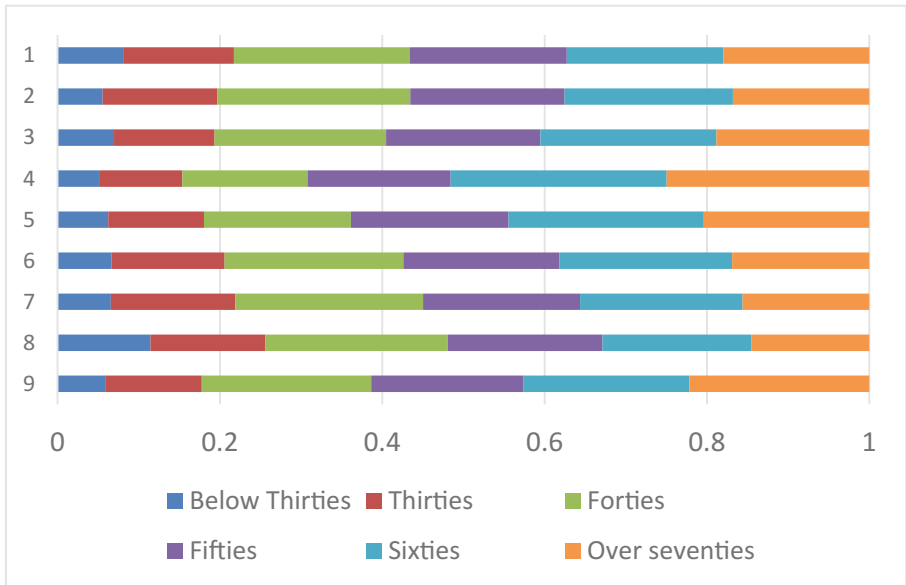**Fig. 5.** Ratio of male and female per cluster

Fig. 6. Ratio of age per cluster

From the Fig. 8 in the ratio of male to female, the cluster with the lowest female ratio is cluster 1, the highest cluster is cluster 9. In the ratio of unmarried and married, the cluster with the lowest marriage ratio is cluster 1, the highest cluster is cluster 2, 7. However, there was not much difference between the clusters in either case.
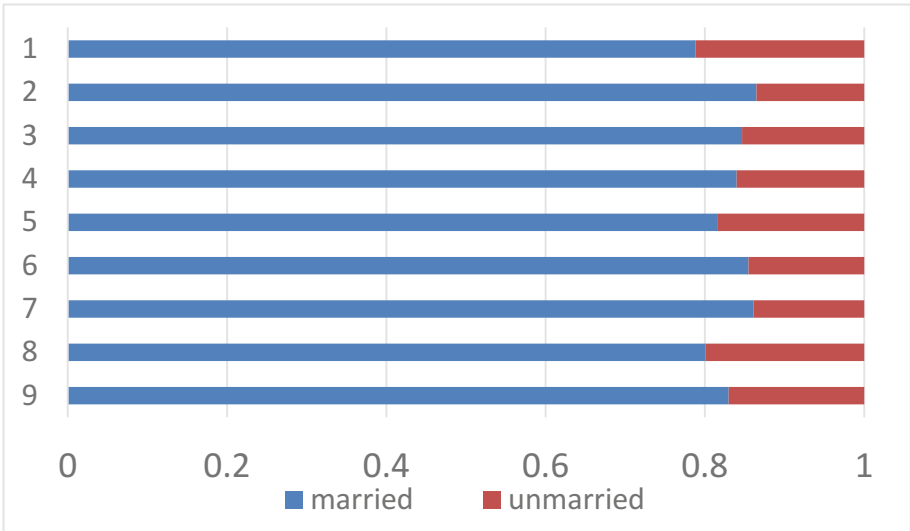

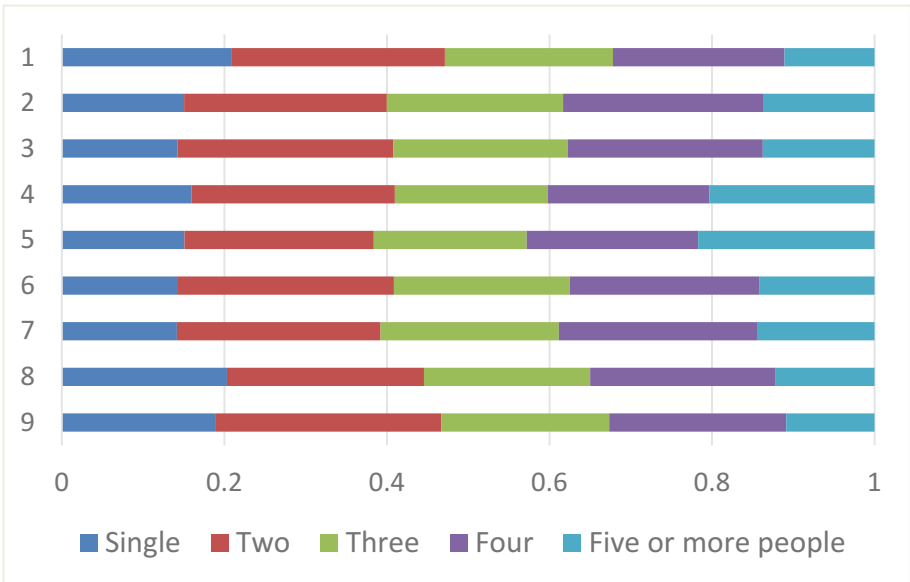**Fig. 7.** Ratio of unmarried and married per cluster

Fig. 8. Ratio of the number of household members per cluster

On the other hand, the number of households and age differed among the clusters. In the ratio of the number of household members, the proportion of "1 person" in clusters 1, 8, 9 is high, while in the clusters 4, 5, the ratio of "5 or more people" is high.

In this way, it was found that there is a difference in characteristics of customers for each cluster (Fig. 7).

5.2 Classification of Customers and Evaluation of Purchasing Behavior

We performed a decile analysis for each cluster and classified our customers as "general customers" and "good customers."

Then, we analyzed purchasing behavior such as aggregation of product categories with high purchase price.

Table 4 shows the number of goods per purchasing opportunity for general customers and good customers is described for each cluster.

Table 4. Number of items per purchase average

Cluster	General customers	Good customers
1	7.0	10.0
2	7.6	10.9
3	6.9	9.7
4	9.0	12.8
5	9.4	12.5
6	8.4	11.8
7	8.2	10.9
8	7.7	11.5
9	5.9	9.0

From Tables 5, 6, 7 and 8 show the top five items of the item category with high purchase price are listed. In this paper, we described only the results of cluster 1 and cluster 4.

The number of items per purchase is higher for good customers than for general customers. Also, it can be seen that there is a difference in the number of items per purchasing opportunity for each cluster.

Table 5. Top 5 items of category in cluster 1 (general customers)

Category name	Total purchase price (ten thousand yen)
Frozen boiled rice	2,029
Sushi	1,833
Brand pork	1,714
Yogurt	1,547
Milk	1,417

Table 6. Top 5 items of category in cluster 1 (good customers)

Category name	Total purchase price (ten thousand yen)
Brand pork	7,602
Japanese beef	7,288
Frozen boiled rice	5,289
Yogurt	5,190
Sushi	4,718

Table 7. Top 5 items of category in cluster 4 (general customers)

Category name	Total purchase price (ten thousand yen)
Sushi	1,998
Japanese beef	1,543
Brand pork	1,005
Frozen chicken	931
Bread	916

Table 8. Top 5 items of category in cluster 4 (good customers)

Category name	Total purchase price (ten thousand yen)
Special Japanese beef	6,735
Brand pork	4,829
Sushi	4,409
Frozen chicken	4,047
Pork	3,472

- Cluster 1

For general customers, merchandise categories that do not require cooking such as ready to eat sushi and frozen boiled rice are higher in purchase price, whereas good customers are higher price of merchandise category requiring cooking such as brand pork and Japanese beef.

Although not listed in the table, beer and the third beer (malt-free beer like alcoholic beverage) were included at the top of purchase price for both general customers and good customers.

- Cluster 4

There was almost no difference between categories where ordinary customers' purchase price was high and categories with high purchase price of superior customers.

5.3   Association Analysis

Using the purchase history from 2015/03/01 to 2016/02/29, we conducted association analysis with basket ID as the key for general customers and good customers in all clusters.

At this time, we extracted the association rule with support: greater than or equal 0.1%, Lift: greater than or equal 1%, rule length: 2 as each threshold (Table 9).

Table 9. Top 5 items of category in cluster 4 (general customers)

Cluster	General customers	Good customers
1	52	272
2	72	458
3	52	270
4	128	792
5	164	658
6	108	564
7	90	558
8	90	430
9	8	170

There were many association rules are extracted for general customers, even for good customers. The clusters with number of rules to be extracted are clusters 4, 5 and 6, and clusters with few rules to be extracted are clusters 1, 3 and 9.

Tables 10, 11, 12 and 13 describe characteristic rules from the extracted association rules. In this study, we described only the results of cluster 1 and cluster 4.

It can be seen that there is a difference between general customers and good customers, we did not find much difference between clusters.

Table 10. The extracted association rule in cluster 1 (general customers)

Antecedent	Consequent	Confidence	Support	Lift
Japanese style confectionery bread	Western style confectionery bread	35.4	2.1	3.3
Banana	Yogurt	20.6	1.2	2.0
Fermented soybeans	Yogurt	20.1	1.3	1.9
Chocolate	Packaged snack	18.0	1.0	2.8
Western style confectionery bread	Yogurt	13.1	1.4	1.3

Table 11. The extracted association rule in cluster 1 (good customers)

Antecedent	Consequent	Confidence	Support	Lift
Chinese cabbage	Tofu	36.9	1.3	2.5
Enoki mushroom	Tofu	36.5	1.6	2.5
Potato	Onion	35.2	1.6	5.8
Frying	Tofu	34.8	1.4	2.4
Carrot	Onion	29.2	1.5	4.8

Table 12. The extracted association rule in cluster 4 (general customers)

Antecedent	Consequent	Confidence	Support	Lift
Japanese style confectionery bread	Western style confectionery bread	35.9	2.4	3.7
Enoki mushroom	Blunder	32.6	1.3	5.8
Chinese noodle	Bean sprouts	30.1	1.3	4.4
Tempura	Tofu	28.6	1.3	2.2
Deep-fried chicken	Salad	18.7	1.2	2.9

Table 13. The extracted association rule in cluster 4 (good customers)

Antecedent	Consequent	Confidence	Support	Lift
Frying	Tofu	51.1	3.8	2.3
Enoki mushroom	Tofu	47.2	2.9	2.2
Chinese cabbage	Tofu	46.2	1.1	2.1
Blunder	Tofu	44.3	4.4	2.0
Fermented soybeans	Tofu	43.6	4.0	2.0

6   Discussions

As the customer characteristic of Clusters 1 and 8, the number of households is one person, the unmarried rate is slightly high, and the proportion of elderly people is low. Clusters 1 and 8 are considered to be clusters where single households in 30 s to 40 s are more than other clusters. On the other hand, as a feature of purchasing behavior, cluster 1 includes beer and third beer in upper category of purchase price. It seems that this is because customers of single households are using stores in Cluster 1, which is centered on small stores, to buy alcoholic drink.

As the customer characteristic of suburban stores in clusters 2, 3, 6 and 7, the ratio of 2 to 4 people in household composition ratio is 70%, that are higher than other clusters, while the ratio of one or five people is low. From these facts, it is speculated that customers are mainly housewives of nuclear families living in the suburbs.

As a characteristic of purchasing behavior, clusters 2, 6, and 7 are such that the number of purchased goods per purchasing opportunity is large and the number of extracted association rules is also large. In Cluster 3, the number of products purchased per purchasing opportunity is small, but this is probably because the average store area is small and the number of products is small. As association rule, typical characteristic was not found. In the clusters 4 and 5 in the mountainous area, the number of households was 5 or more, and the ratio of households in the 60 s and over 70 s was higher in the ages. From these facts, clusters 4 and 5 are considered to be clusters of elderly people couple families who live in mountainous areas compared with other clusters.

The characteristic of purchasing behavior is that the purchase price of sushi is high, the number of purchased goods per purchasing opportunity is large, and the number of extracted association rules is large.

Also, since it is a cluster with an average parking number of 100 or more, it is thought that there are many customers who visit by car and buy many items at a time.

As a feature of purchasing behavior common to all clusters, general customers have higher product categories that do not require cooking, whereas good customers tend to have higher-ranking product categories requiring cooking. In addition, as a result of association analysis, rules such as "chocolate - snack" and "banana - yogurt" are found for general customers, and rules such as "potatoes - onions" and "wooden mushrooms - tofu" are found for good customers.

For this reason, we think that there are differences in the purpose of using super-markets for general customers and good customers, and it is necessary to propose appropriate measures for each.

7 Conclusion and Future Works

Study, first, we classified stores by cluster analysis using store causal data. Second, focusing on each classified cluster, we aggregate customer characteristics such as age structure and household composition and classified customers. In addition, we performed association analysis on purchasing behaviors and clarified concurrent selling commodities for each cluster. Finally, we clarified customer characteristics and purchasing behavior for each store causal from these analysis results.

Own analysis revealed that there are differences in customer characteristics and purchasing behavior depending on the characteristics of each store, such as the sales floor area and the population within the trading area. By using this result, it is possible to propose marketing measures unique to that store according to the surrounding environment, store size, etc., which were not taken into account in analysis using only the purchase history. In order to classify the stores more accurately, it is necessary to consider a method for the commercial areas of each store, taking into account competing stores and the like. In addition, it is thought that more useful suggestions can be obtained by looking at changes in purchasing trends such as the seasons of each cluster and considering customers using many stores.

References

1. ATKearney. https://www.atkearney.com/consumer-products-retail/ideas-insights/featured-article/-/asset_publisher/KQNW4F0xInID/content/online-retail-is-front-and-center-in-the-quest-for-growth/10192
2. Romesburg, C.: Cluster Analysis for Researchers, pp. 133–135. Lulu.com
3. Decile Analysis. http://www.totalcustomeranalytics.com/decile_analysis.htm
4. Nagasawa, T., Yamagishi, A., Yokoyama, S.: Analysis of consumer purchase behavior using supermarket sales data and regional information. Commun. Jpn. Ind. Manag. Assoc. 25(3), 158–163 (2015)
5. Namatame, T., Suyama, N.: Weather effects on consumer behavior in retailing - an analysis by using a supermarket's POS data. Bull. Inst. Commer. 41(8), 1–29 (2010)
6. Sato, M., Kato, E., Matsuda, Y.: Research of FSP analysis in food supermarket. UNISYS Technol. Rev. 25(3), 330–338 (2005)
7. Annie, L.C., Kumar, A.: Market basket analysis for a supermarket based on frequent itemset mining. Int. J. Comput. Sci. 9(5), 257–264 (2012)