International Journal of Institutional Research and Management
International Institute of Applied Informatics
2017, Vol.1, No.1, P.1 - 20

Development of Recommendation Engines for Enhancing Sales of DIY (Do It Yourself) Stores: Vertical Approach vs. Horizontal Approach

Xinlong Hu [*], Shan Yu [*],
Masaki Shobu [*], Ushio Sumita [†]

Abstract

In Japan, the popularity of DIY stores has been growing rapidly. In comparison with typical large retail chain stores, DIY stores have a wider variety of products mostly with lower prices and broader store spaces. In many cases, they are located in suburban areas with huge parking lots. Because of these unique features, customer behaviors for DIY stores could be quite different from those for ordinary large retail chain stores, and some special attentions may be needed for sales promotion. The purpose of this paper is to establish a framework for developing recommendation engines for DIY stores so as to enhance their sales. More specifically, useful recommendation rules are derived from two different perspectives: a vertical approach from a point of view of pairs of products across product categories with significant sales contributions, and a horizontal approach focusing on excellent customers who are ranked in a top segment in terms of both the purchasing amount of money and the purchasing volume of products. Assuming that certain marketing campaigns are conducted along the derived association rules, a computational procedure is developed for assessing the economic impact of each of such recommendation rules.

Keywords: DIY stores, recommendation engine, vertical and horizontal approach, association rules, economic impact.

1 Introduction

In order to facilitate individual marketing decisions of customers in choosing what they really want out of huge options available through the Internet, recommender systems have been employed in e-commerce applications. The extensive literature exists concerning how such recommender systems may be developed. In [1], for example, several techniques are presented for analyzing large-scale purchase and preference data, where data mining, nearest neighbor collaborative filtering and dimensionality reduction techniques are combined together to establish recommendation rules for customers. Performances of major association rule algorithms are compared in [3] using real data. Reference [6] attempts to develop emotion-based music recommendation rules by association discovery from film music. A hybrid approach is proposed in [7] by exploiting advantages of a weighted RFM (Recency, Frequency and Monetary) method together with a preference based collaborative filtering method. Recently, the problems of security and protection of personal information associated with recommender systems have been also drawing attention of many researchers, see e.g. [8] and [9].

---

[*] Department of Information Excellence, LIXIL Corporation, Tokyo, JAPAN
[†] Graduate School of Business Administration, Keio University, Kanagawa, JAPAN

In parallel with the line of research discussed above, this paper addresses itself to the problem of how to develop recommendation engines for enhancing sales of DIY stores. While this problem may not necessarily involve on-line shopping, consumers at DIY stores face challenges similar to those for on-line shopping in that DIY stores offer a variety of products ranging from daily necessities to furniture and pets over vast store spaces. Accordingly, some promotional guidance may help customers to facilitate their purchasing decisions. Providing efficient recommendation engines useful for customers would also help DIY stores to increase cross-selling, enhance customer loyalty and improve customer satisfaction.

The structure of this paper is as follows. Section 2 provides a succinct summary of data to be employed for the study. The data have been obtained from a Japanese company operating a chain of DIY stores across Japan. Based on a preliminary analysis, a store and two sales areas of significant importance for the company are chosen for further study, so as to contain the computational burden arising from the massive data. However, the methodological approach proposed in this paper would be applicable to other sales areas as well as other DIY stores. In Section 3, a vertical approach is described for establishing recommendation rules. For the pair of sales areas selected in the previous section, a pair of large categories of products would be selected, followed by identification of a pair of medium categories within the selected pair. This procedure is repeated until small categories of products of importance are identified. Then extracted is the class of customers who contribute to the sales of the products in the selected small categories. Focusing on those small categories and the associated customer class, recommendation rules are established. Section 4 is dedicated to development of a horizontal approach, where we first identify a class of customers of importance in terms of both the purchasing amount of money and the purchasing volume of products over the selected pair of sales areas, and then recommendation rules are derived across pairs of small categories of products within the selected pair of sales areas among the class of customers. In Section 5, the vertical approach is compared with the horizontal approach by examining the recommendation rules resulting from each approach. Section 6 discusses a management scheme for facilitating the decision of whether or not a currently adopted recommendation rule should be continued for the next month. A computational procedure is established in Section 7 for assessing the economic impact of any recommendation rule derived from either the vertical approach or the horizontal approach. Some final remarks are given in Section 8.

2 Data Description and Selection of a Store and two Sales Areas for the Study

In this paper, we employ real data obtained from a Japanese DIY store company, concerning all of the stores of the company across Japan in the fiscal year 2014. This data would be employed to choose a store of importance for the company, representing all other stores for establishing recommendation engines. Since a new POS system was introduced in the middle of May 2015, however, more recent transaction data will be used, covering the period of June 1st through November 30th 2015 at the chosen store, so as to select a pair of two sales areas based on their contributions to the amount of money purchased and the volume of products sold, as well as “simultaneous purchasing” across the two sales areas. Furthermore, further analysis would be conducted toward development of association rules for sales recommendation. It is worth noting that the approach proposed here would be applicable to any other DIY store.

TABLE 2.1 Definition of Store Size

	Store Areas (1000㎡)	Gross Sales (￥billion)
Small	Less than 5	Less than 2
Medium	5〜13.2	2〜4
Large	More than 13.2	More than 4

TABLE 2.2 Compositions of Stores

		Gross Sales			Total
		Small	Medium	Large	
Areas	Small	38	1	0	15
	Medium	10	9	1	20
	Large	0	11	4	39
Total		48	21	5	74

The company under consideration well represents a typical DIY store company in Japan in terms of the number of stores, geographical areas covered by those stores, ranges of products sold and sales volume. More specifically, the DIY store company has 74 stores across Japan, where the three classes Small, Medium, and Large for store areas and those for gross sales are defined as in TABLE 2.1. According to the three classes for store areas and the three classes for gross sales, these stores can be classified into groups, as shown in TABLE 2.2.

In terms of products, they are classified according to the hierarchical structure of 5 layers, as depicted in Figure 2.1, starting with 16 sales areas followed by 56 large categories, 347 medium categories, 1458 small categories, and finally 522,214 products at the bottom.

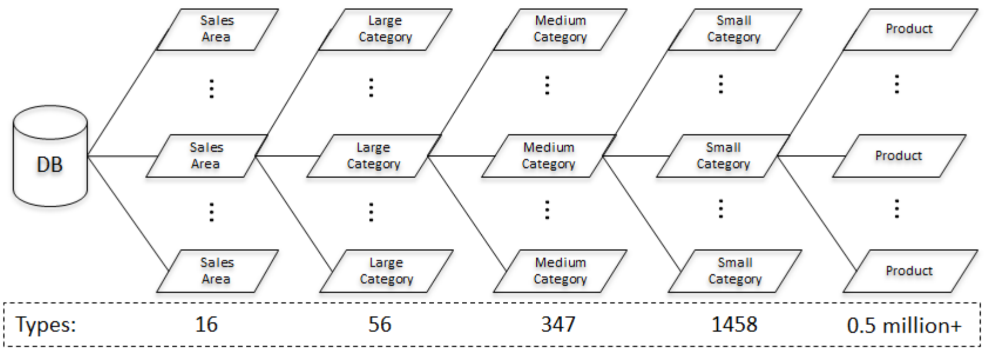


Figure 2.1 Hierarchical Structures of Products

In order to develop recommendation engines for sales promotion, two different approaches may be considered. The first approach is called a vertical approach, where association rules across different products or product categories are identified based on analyses of "simultaneous purchasing." This means that the sales records of different products or product categories with high correlations within a specified time period are detected. The time period for determining "simultaneous purchasing" may vary depending on products or product categories. The term "vertical" is used because the association analyses progress along the hierarchical structure depicted in Figure 2.1 from the top toward the bottom. This approach is reasonable because the number of possible combinations of products or lower level product categories is too large to search for effective association rules directly, and the amount of time needed for the direct search is simply overwhelming.

The second approach is called a horizontal approach, where a class of customers is specified first, and then association rules across different products or product categories are identified as for the vertical approach but within the customer class. The term "horizontal" means that association rules established through the horizontal approach may involve products or product categories without any regard to the hierarchical structure in Figure 2.1.

The first step for our study is to select a store of significant business importance and restrict our analysis to this store so as to contain the potential difficulty associated with the massive data volume. For this purpose, all stores are sorted first in descending order in terms of the gross sales as well as the number of products purchased. The cumulative sum of each of the two is then computed from the top, along with the corresponding percentage. By decomposing all stores into 100 groups where separations are made by drawing a line at each incremental 10 percent point along the two axes, the two dimensional decile matrix for stores can be constructed, as shown in TABLE 2.3.

TABLE 2.3 Two Dimensional Decile Matrix for Stores

		Gross Sales									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number of Products Purchased	10%	Store 1	0	0	0	0	0	0	0	0	0
	20%	0	Store 2 Store 3	1	0	0	0	0	0	0	0
	30%	0	0	2	2	0	0	0	0	0	0
	40%	0	0	1	2	1	0	0	0	0	0
	50%	0	0	0	1	3	0	0	0	0	0
	60%	0	0	0	0	1	4	2	0	0	0
	70%	0	0	0	0	0	2	2	3	1	0
	80%	0	0	0	0	0	0	2	6	2	0
	90%	0	0	0	0	0	0	0	1	11	3
	100%	0	0	0	0	0	0	0	0	1	28

It can be readily seen that three stores contribute to 20% of the sum of the gross sales over all the stores, as well as that of the number of products purchased over all the stores. In particular, the contribution of Store 1 alone amounts to 10% of each sum. Because of this reason, we exclusively focus on Store 1 throughout the remainder of this paper. However, the methodological approach proposed in this paper can be applicable to any other store of interest.

In order to derive some recommendation rules explicitly for Store 1, we work on real data which consist of customer purchasing records during the period June 1st through November 30th, 2015. The number of customers who purchased at least one product during the data period amounts to 39,477. A pair of sales areas is then selected based on their contributions to the amount of money and the volume of products purchased for further analysis. TABLE 2.4 depicts the two dimensional decile matrix for sales areas at Store 1. One sees that two sales areas combined together account for 40% of both the gross sales and the number of products purchased.

TABLE 2.4 Two Dimensional Decile Matrix for Sales Areas

		Gross Sales (About ¥ 3 billion)									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number of Products Purchased (About ¥6 million)	10%	0	0	0	0	0	0	0	0	0	0
	20%	0	0	0	0	0	0	0	0	0	0
	30%	0	0	0	1	0	0	0	0	0	0
	40%	0	1	0	0	0	0	0	0	0	0
	50%	0	0	0	0	0	0	0	0	0	0
	60%	0	0	0	0	1	0	0	0	0	0
	70%	0	0	0	0	0	0	0	0	1	0
	80%	0	0	0	0	0	1	1	0	0	0
	90%	0	0	0	0	0	0	0	0	1	0
	100%	0	0	0	0	0	0	0	2	0	7

We next examine interrelations among different sales areas in terms of how customers purchased products across such sales areas over the 6 month data period. In other words, "simultaneous purchasing" over two different sales areas is considered using 6 months as the time unit. Given a sales area Z, let N and $N(Z)$ be defined by

N : the number of customers who purchased at least once at Store 1 over the data period ; and

$N(Z)$: the number of customers who purchased at least one of the products which belong to Z at Store 1 over the data period.

For two different sales areas X and Y, we also define $Supp(X,Y)$ and $Conf(X,Y)$ by

$$Supp(X,Y) = \frac{N(X \cap Y)}{N} \; ; \quad Conf(X,Y) = \frac{N(X \cap Y)}{N(X)} \; . \tag{1}$$

It should be noted that, higher the value of $Supp(X,Y)$ is, more likely that "simultaneous purchasing" exists between X and Y. Furthermore, if $Conf(X,Y) > Conf(Y,X)$, it is more likely that the statement "if a customer purchases X, then he/she also purchases Y" (denoted by $X \rightarrow Y$, hereafter) holds true, rather than $Y \rightarrow X$.

In Figure 2.2, the interrelations across different sales areas are depicted in terms of $Supp$ and $Conf$. The thickness of the arrow from X to Y indicates the strength of $Supp(X,Y)$, which is written as S in the middle of the three numbers on the arrow. The first number $C1$ is the value of $Conf(Y|X)$ and the last number $C2$ is that of $Conf(Y|X)$ with $C1 > C2$.

By examining Figure 2.2 closely together with TABLE 2.4, one can identify important interrelations of interest for further study. In this paper, as such an example, we focus on Household Goods → Daily Necessities ($S = 0.54 \; and \; C1 = 0.9$) .

The two sales areas involved above are located in the upper left corner in TABLE 2.4, meaning that they are important in terms of the gross sales as well as the number of products purchased.

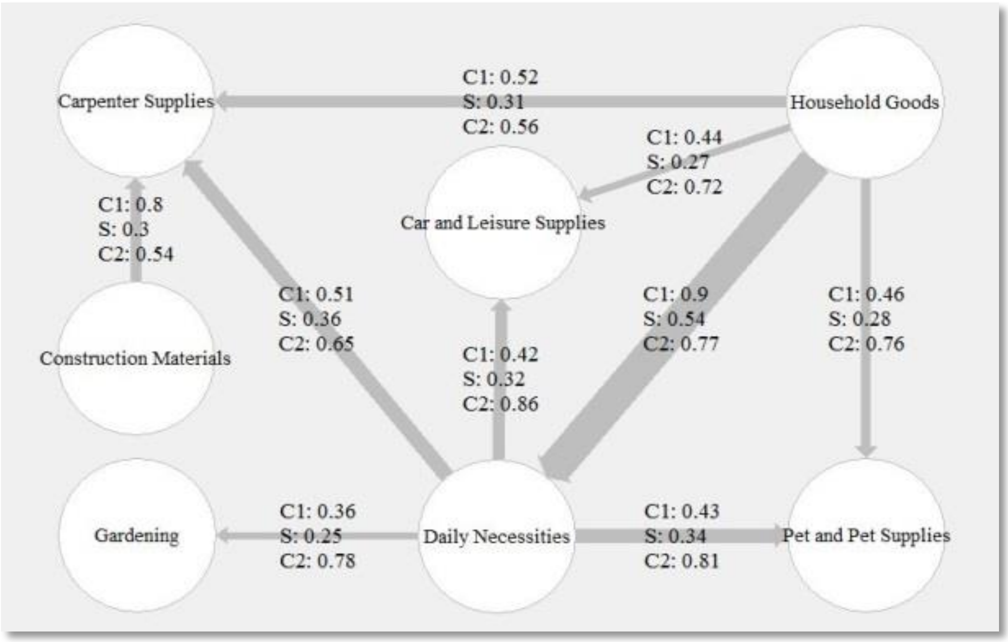


Figure 2.2 Interrelations among Different Sales Areas
(Threshold: Supp ≥ 0.1; Conf ≥ 0.1)

Clearly, the time unit of 6 months is too long for developing any realistic recommendation rule for DIY stores. However, the analysis so far is still important in that if "simultaneous purchasing" does not exist between two sales areas with this time unit, then it will not exist with any shorter time unit. The time unit will be shortened as the vertical approach progresses from sales areas to medium and small categories, as we will see.

3 Vertical Approach

In this section, we select small categories and the associated customer class of interest for establishing association rules based on the vertical approach. Some of the resulting recommendation rules are also presented.

Let $HD\text{-}SAG$ (Sales Area Group) be the set of sales areas that belong to Household Goods and Daily Necessities. We define $HD\text{-}MCG$ (Medium Category Group) similarly for medium categories. The purpose of this section is first to select medium categories in $HD\text{-}MCG$ together with an associated class of customers of importance for further study. Within the selected medium categories and the customer class, we next select a set of small categories as well as an associated customer class, which will provide a basis for deriving actual recommendation rules.

In TABLE 3.1 below, a number in a cell indicates the number of medium categories that contribute to the corresponding percentages in terms of the gross sales and the number of products purchased for $HD\text{-}SAG$. One sees that 12 medium categories out of 37, or 32% of the total medium categories in $HD\text{-}SAG$, collectively account for 80% of both the gross sales and the number of products purchased for $HD\text{-}SAG$. These 12 medium categories are listed in TABLE 3.2. Interrelations among the 12 medium categories are illustrated in Figure 3.1, demonstrating that all of them are worth for consideration for further study.

TABLE 3.1 Two Dimensional Cumulative Decile Matrix for Medium Categories in HD-SAG

HD-SAG		Gross Sales									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number of Products Purchased	10%	0	0	0	0	0	0	0	0	0	0
	20%	0	1	1	1	1	1	1	1	1	1
	30%	0	1	1	2	2	2	2	2	2	2
	40%	0	1	1	2	3	4	4	4	4	4
	50%	0	1	2	3	4	5	5	6	6	6
	60%	0	1	2	4	5	6	6	8	8	8
	70%	0	1	2	4	5	6	9	11	12	12
	80%	0	1	2	4	5	7	10	12	15	15
	90%	0	1	2	4	5	7	10	13	16	21
	100%	0	1	2	4	5	7	10	13	19	37

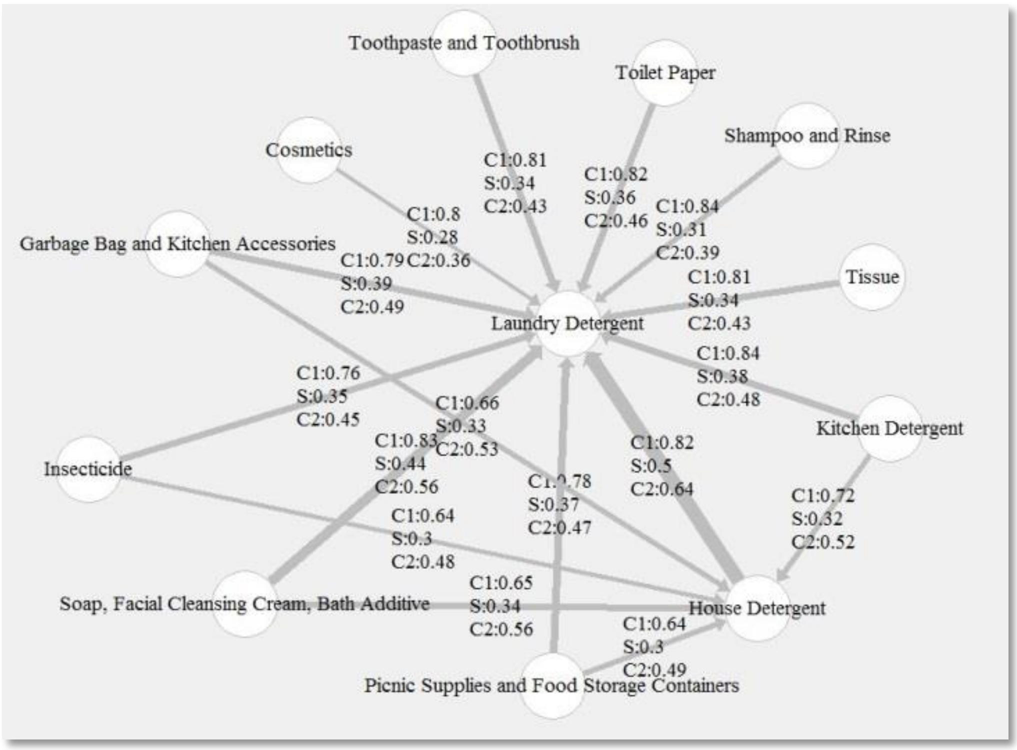


Figure 3.1 Interrelations among 12 Selected Medium Categories
(Threshold: $Supp \geq 0.25; Conf \geq 0.1$)

TABLE 3.2 Selected Medium Categories in HD-SAG

I	House Detergent
II	Laundry Detergent
III	Soap, Facial Cleansing Cream, Bath Additive
IV	Garbage Bag and Kitchen Accessories
V	Kitchen Detergent
VI	Picnic Supplies and Food Storage Containers
VII	Toilet Paper
VIII	Insecticide
IX	Toothpaste and Toothbrush
X	Tissue
XI	Shampoo and Rinse
XII	Cosmetics

Let $HD\text{-}MC\text{-}SEL$ be the set of the selected medium categories listed in TABLE 3.2. In parallel with TABLE 3.1 but for customers rather than medium categories, we construct TABLE 3.3 by restricting the set of products for computing the gross sales and the number of products purchased to those products within $HD\text{-}MC\text{-}SEL$. It can be seen that 8,947 customers out of 39,477, or 23% of the entire customers, collectively contribute to 80% of both the gross sales and the number of products purchased for $HD\text{-}MC\text{-}SEL$. The resulting set of customers is denoted by $HD\text{-}MC\text{-}SEL\text{-}C\text{-}80$. We now focus on $HD\text{-}MC\text{-}SEL$ and $HD\text{-}MC\text{-}SEL\text{-}C\text{-}80$ so as to further narrow them down to a set of small categories and an associated customer class.

TABLE 3.3 Two Dimensional Cumulative Decile Matrix for Customers Associated with HD-MC-SEL

HD-MC-SEL		Gross Sales									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number of Products Purchased	10%	2	2	2	2	2	2	2	2	2	2
	20%	99	277	307	314	321	324	324	324	324	324
	30%	115	569	880	991	1035	1054	1060	1064	1066	1067
	40%	117	679	1318	1731	1950	2025	2060	2080	2088	2093
	50%	118	706	1513	2301	2870	3169	3310	3380	3407	3417
	60%	119	715	1566	2590	3545	4303	4753	4966	5052	5096
	70%	119	722	1590	2703	3943	5236	6217	6872	7146	7250
	80%	120	727	1600	2734	4115	5790	7446	8947	9815	10120
	90%	121	732	1610	2757	4177	5964	8153	10811	13199	14333
	100%	122	736	1626	2780	4238	6064	8398	11515	16066	39477

Let $HD\text{-}MC\text{-}SEL\text{-}SC$ be the set of small categories within $HD\text{-}MC\text{-}SEL$. TABLE 3.4 below is constructed in a manner similar to TABLE 3.1. One sees that $HD\text{-}MC\text{-}SEL\text{-}SC$ contains 69 small categories and the contributions of only 21 small categories, or 30% of the categories in $HD\text{-}MC\text{-}SEL\text{-}SC$, account for 80% of both the gross sales and the number of products purchased for $HD\text{-}MC\text{-}SEL$. These small categories are listed in TABLE 3.5.

TABLE 3.4 Two Dimensional Cumulative Decile Matrix for Small Categories within HD-MC-SEL

HD-MC-SEL		Gross Sales									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number of Products Purchased	10%	0	0	0	0	0	0	0	0	0	0
	20%	1	1	1	1	1	2	2	2	2	2
	30%	1	2	2	3	3	4	4	4	4	4
	40%	1	3	3	4	4	5	5	6	6	6
	50%	1	3	4	5	5	7	7	9	9	9
	60%	1	3	5	7	9	11	11	13	13	13
	70%	1	3	5	7	11	14	14	17	19	19
	80%	1	3	5	7	11	14	17	21	27	27
	90%	1	3	5	7	11	15	20	26	34	38
	100%	1	3	5	7	11	15	20	27	36	69

TABLE 3.5 Selected Small Categories within HD-MC-SEL

I	Thermos
II	Cling Film and Foil
III	Garbage Bag
IV	Bathtub Cleaner
V	Toilet Cleaner
VI	Wet Tissue
VII	Powder Detergent
VIII	Liquid Detergent
IX	Laundry Aids
X	Fabric Softener
XI	Bleach Cleaner
XII	Dishwashing Detergent
XIII	Shampoo
XIV	Toothpaste
XV	Soap, Facial Cleansing Cream, Bath Additive
XVI	Bath Additive
XVII	Boxed Tissue
XVIII	Recycled Toilet Paper
XIX	Toilet Paper
XXX	Mosquito Liquid Electronic Repellent
XXI	Repellents

Let $HD\text{-}SC\text{-}SEL$ be the set of the selected small categories listed in TABLE 3.5. We construct TABLE 3.6 as for TABLE 3.3, where the set of products for computing the gross sales and the number of products purchased is restricted to those products within $HD\text{-}SC\text{-}SEL$. One observes that 7,804 customers, or 20% of the entire customers, contribute to 80% of both the gross sales and the number of products purchased for $HD\text{-}SC\text{-}SEL$. The resulting set of customers is denoted by $HD\text{-}SC\text{-}SEL\text{-}C$.

TABLE 3.6 Two Dimensional Cumulative Decile Matrix for Customers Associated with HD-SC-SEL

HD-SC-SEL		Gross Sales									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number of Products Purchased	10%	1	1	1	1	1	1	1	1	1	1
	20%	53	76	81	81	83	83	84	84	84	84
	30%	111	429	565	613	634	644	647	649	655	655
	40%	117	587	998	1290	1419	1486	1502	1516	1529	1531
	50%	122	643	1271	1865	2282	2504	2585	2641	2672	2688
	60%	123	664	1426	2268	3036	3637	3874	4072	4133	4171
	70%	124	677	1477	2437	3490	4518	5254	5824	5995	6094
	80%	125	684	1493	2512	3738	5138	6547	7804	8418	8662
	90%	126	687	1503	2549	3852	5450	7412	9451	11542	12463
	100%	132	706	1530	2595	3934	5621	7779	10649	14879	39477

Each of the 21 small categories in TABLE 3.5 contains many similar products. A small category "Cling Film and Foil," for example, contains two products "X Corp Cling Film 22×50m" and "X Corp Cling Film 30×50m." Similarly, a small category "Shampoo" has two products "Y Corp Rich Shampoo for Refill" and " Y Corp Rich Conditioner for Refill." Association rules among such similar products are of little use. Accordingly, we restrict ourselves to small categories for deriving association rules, that is, we concentrate on deriving association rules among the small category class $HD\text{-}SC\text{-}SEL$ in TABLE 3.5 with focus on the customer class $HD\text{-}SC\text{-}SEL\text{-}C\text{-}80$ having 7,804 customers.

Focusing on the small categories and the associated customer class derived above, we are now in a position to establish association rules $X \rightarrow Y$ through the procedure discussed in Section 2. More specifically, let N be the number of customers under consideration and let *N(Z)* be the number of customers who purchased at least one of the products that belong to the small category Z. One can then define $Supp(X, Y)$ and $Conf(X, Y)$ as in (2.1). Some of the resulting recommendation rules with $Supp(X, Y) \geq 0.01$ and $Conf(X, Y) \geq 0.01$ are listed in TABLE 3.7 in the order of $Conf(X, Y)$.

TABLE 3.7 Association Rules X→Y with Supp ≥ 0.01; Conf ≥ 0.01

Rules of Vertical Approach				
X	Y	Supp	Conf	Lift
Bleach Cleaner	Fabric Softener	0.187	0.722	1.269
Liquid Detergent	Fabric Softener	0.365	0.678	1.191
Bleach Cleaner	Liquid Detergent	0.172	0.663	1.232
Fabric Softener	Liquid Detergent	0.365	0.641	1.191
Bathtub Cleaner	Fabric Softener	0.174	0.621	1.092
Dishwashing Detergent	Fabric Softener	0.232	0.620	1.090
Laundry Aids	Fabric Softener	0.097	0.619	1.087
Shampoo	Fabric Softener	0.218	0.619	1.087
Toothpaste	Fabric Softener	0.135	0.611	1.074
Soap, Facial Cleansing Cream, Bath Additive	Fabric Softener	0.277	0.604	1.062
Bath Additive	Fabric Softener	0.076	0.604	1.061
Toilet Cleaner	Fabric Softener	0.121	0.603	1.060
Laundry Aids	Liquid Detergent	0.094	0.602	1.119
Wet Tissue	Fabric Softener	0.152	0.599	1.052
Cling Film and Foil	Fabric Softener	0.178	0.590	1.036

4 Horizontal Approach

The first step in the horizontal approach is to identify a class of customers who significantly contribute to both the gross sales and the number of products purchased within the two sales areas selected in Section 2. Among such customers in this class, a class of small categories of key importance across the two sales areas would be identified. More specifically, as for TABLE 3.4 and TABLE 3.6 in Section 3, one can obtain TABLE 4.1 and TABLE 4.2, where each cell shows the cumulative number of customers in TABLE 4.1 and the cumulative number of small categories in TABLE 4.2, contributing to the two percentiles specified by the cell. One finds that 9640 customers (or 24.4% of 39477 customers) contributed to 80% of both the gross sales and the number of products purchased in TABLE 4.1, while 46 small categories (or 24.6% of 187 small categories) in TABLE 4.2.

TABLE 4.1 Two Dimensional Cumulative Decile Matrix for Customers Associated with Selected Two Sales Areas

HD-HORI-C		Gross Sales									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number of Products Purchased	10%	4	4	4	4	4	4	4	4	4	4
	20%	107	334	379	396	408	418	422	424	424	429
	30%	118	619	960	1095	1156	1191	1216	1230	1238	1247
	40%	119	705	1397	1864	2115	2211	2275	2318	2342	2364
	50%	120	716	1562	2434	3072	3372	3571	3691	3762	3810
	60%	121	722	1596	2679	3777	4624	5114	5409	5568	5655
	70%	121	727	1607	2741	4091	5583	6674	7425	7801	8006
	80%	122	730	1614	2759	4202	5986	7910	9640	10645	11126
	90%	122	735	1619	2771	4224	6044	8360	11280	14261	15689
	100%	122	736	1626	2780	4238	6064	8398	11515	16066	39477

TABLE 4.2 Two Dimensional Cumulative Decile Matrix for Small Categories within Selected Two Sales Areas

HD-HORI-S		Gross Sales									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number of Products Purchased	10%	0	0	0	0	1	1	1	1	1	1
	20%	1	2	2	2	3	3	3	3	3	3
	30%	1	3	4	5	6	6	7	7	7	7
	40%	1	4	6	7	9	9	11	11	11	11
	50%	1	4	7	10	13	14	17	17	18	18
	60%	1	4	7	12	16	18	21	27	28	28
	70%	1	4	7	12	17	22	26	37	40	41
	80%	1	4	7	12	18	25	33	46	56	58
	90%	1	4	7	12	18	25	34	48	67	85
	100%	1	4	7	12	18	25	34	48	73	187

As for the vertical approach, key association rules $X \rightarrow Y$ can now be established by focusing on the all small categories within Selected Two Sales Areas and the associated customer class derived above, with the selection criteria of $Supp(X,Y) \geq 0.01$ and $Conf(X,Y) \geq 0.01$ again. Some of the resulting recommendation rules are listed in TABLE 4.3 in the order of $Conf(X,Y)$.

TABLE 4.3 Association Rules X→Y with

Rules of Horizontal Approach				
X	Y	Supp	Conf	Lift
Bleach Cleaner	Fabric Softener	0.152	0.692	1.426
Wool Washing Detergent	Fabric Softener	0.035	0.671	1.384
Liquid Detergent	Fabric Softener	0.292	0.646	1.333
Children Toothbrush	Fabric Softener	0.020	0.639	1.317
Wool Washing Detergent	Liquid Detergent	0.033	0.630	1.392
Sanitary Items for Night Use	Fabric Softener	0.029	0.628	1.294
Bleach Cleaner	Liquid Detergent	0.138	0.626	1.385
Sanitary Items for Day Use	Fabric Softener	0.057	0.616	1.270
Soap, Facial Cleansing Cream, Bath Additive	Fabric Softener	0.077	0.615	1.269
Children Toothbrush	Liquid Detergent	0.019	0.615	1.359
Fabric Softener	Liquid Detergent	0.292	0.603	1.333
Children Toothbrush	Soap, Facial Cleansing Cream, Bath Additive	0.018	0.601	1.504
Rinse Treatments	Soap, Facial Cleansing Cream, Bath Additive	0.044	0.599	1.497
Soap, Facial Cleansing Cream, Bath Additive	Soap, Facial Cleansing Cream, Bath Additive	0.075	0.596	1.489
Shampoo	Fabric Softener	0.172	0.588	1.213

5 Comparison of Vertical Approach with Horizontal Approach

In this section, we compare association rules derived from the vertical approach with those based on the horizontal approach. In order to facilitate the writing of the comparative analysis, we provide IDs to relevant small categories under consideration. The list of such IDs is given in TABLE 5.1.

TABLE 5.1 List of IDs of Relevant Small Categories

ID	Small Category Name	ID	Small Category Name
S_1	Bleach Cleaner	S_13	Cling Film and Foil
S_2	Liquid Detergent	S_14	Recycled Toilet Paper
S_3	Fabric Softener	S_15	Powder Detergent
S_4	Bathtub Cleaner	S_16	Wool Washing Detergent
S_5	Dishwashing Detergent	S_17	Children Toothbrush
S_6	Laundry Aids	S_18	Sanitary Items for Night Use
S_7	Shampoo	S_19	Sanitary Items for Day Use
S_8	Toothpaste	S_20	Facial Cleansing Cream
S_9	Soap, Facial Cleansing Cream, Bath Additive	S_21	Rinse Treatments
S_10	Bath Additive	S_22	Other Sanitary Items
S_11	Toilet Cleaner	S_23	Diaper
S_12	Wet Tissue	S_24	Garbage Bag

We note from TABLE 3.4 and TABLE 4.1 that 69 small categories are selected through the vertical approach, while 187 small categories are extracted based on the horizontal approach. Furthermore, those small categories that belong to the top 40% in each dimension of the underlying two dimensional cumulative decile matrix are listed for the vertical approach in TABLE 5.2 (a) and for the horizontal approach in TABLE 5.2 (b). One can find that all of the small categories in TABLE 5.2 (a) appear also in TABLE 5.2 (b), where the former small categories are located further to the right of those in TABLE 5.2 (b). This means that association rules for recommendation based on the horizontal approach are likely to contribute more in terms of the gross sales than those derived through the vertical approach. For each month of June through November 2015, the number of association rules for recommendation based on the vertical approach is compared with that associated with the horizontal approach, where the selection criteria are Supp $\geq$ 0.01 and Conf $\geq$ 0.01. It can be seen in TABLE 5.3 that the horizontal approach extracts at least 3.6 times more association rules than the vertical approach.

TABLE 5.2 Small Categories within Top 40% of Two Dimensional Cumulative Decile Matrix

(a) Vertical Approach

Vertical Approach		Gross Sales 10%	20%	30%	40%
Number of Products Purchased	10%			S_3	
	20%				
	30%		S_2		
	40%			S_9	S_7

(b) Horizontal Approach

Horizontal Approach		Gross Sales 10%	20%	30%	40%
Number of Products Purchased	10%				
	20%	S_2 S_3		S_7	
	30%		S_9		S_23
	40%		S_24	S_5	

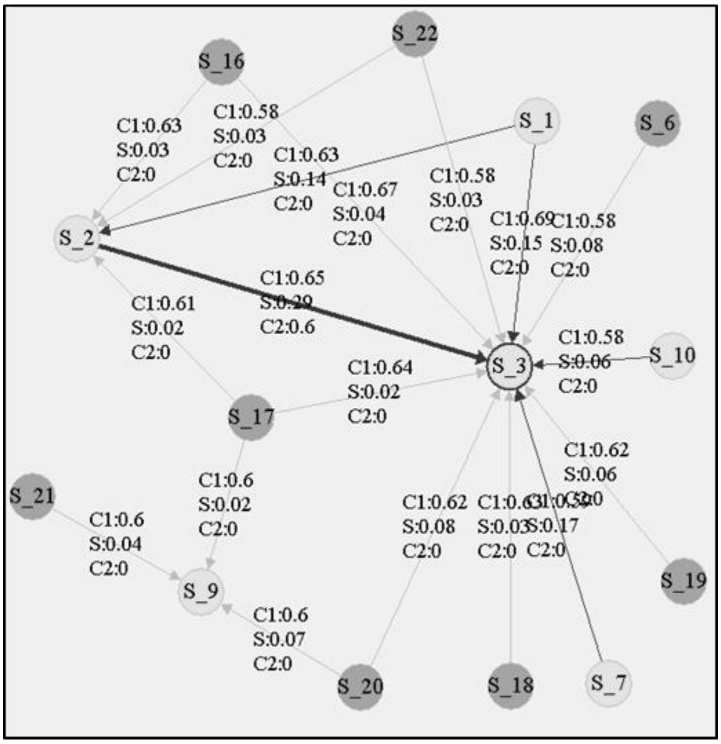
TABLE 5.3 Number of Recommendation Rules Satisfying Supp ≥ 0.01; Conf ≥ 0.01

	Jun	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**
Vertical	392	416	377	482	377	379
Horizontal	2225	2152	2091	1732	2038	2015

Figure 5.1 (a) depicts association rules for recommendation based on the vertical approach, while Figure 5.1 (b) provides the counterpart for the horizontal approach. It should be noted that several strong association rules are common across the two approaches, as shown in TABLE 5.4. It may be worth noting that these common rules are stronger with the horizontal approach than with the vertical approach in terms of both Supp and Conf. Some rules hold true only for one of the two approaches, also shown in TABLE 5.4.

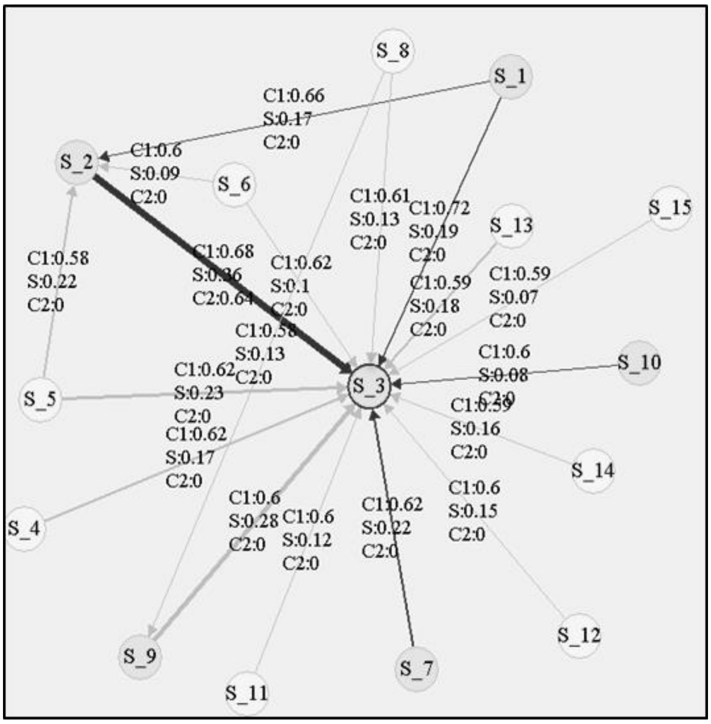
TABLE 5.4 Association Rules Derived by Two Approaches in Figure 5.1

		Vertical Approach		Horizontal Approach	
X	Y	Supp	Conf	Supp	Conf
S_1	S_2	0.14	0.63	0.17	0.66
S_1	S_3	0.15	0.69	0.19	0.72
S_2	S_3	0.29	0.65	0.36	0.68
S_7	S_3	0.17	0.59	0.22	0.62
S_10	S_3	0.06	0.58	0.08	0.6
S_6	S_3	0.08	0.58	-	-
S_16	S_3	0.04	0.67	-	-
S_17	S_3	0.02	0.64	-	-
S_17	S_9	0.02	0.60	-	-
S_17	S_2	0.02	0.61	-	-
S_18	S_3	0.03	0.63	-	-
S_19	S_3	0.06	0.62	-	-
S_21	S_9	0.04	0.60	-	-
S_22	S_3	0.03	0.58	-	-
S_22	S_2	0.03	0.58	-	-
S_4	S_3	-	-	0.17	0.62
S_5	S_3	-	-	0.23	0.62
S_6	S_2	-	-	0.09	0.60
S_8	S_3	-	-	0.13	0.61
S_8	S_9	-	-	0.13	0.58
S_11	S_3	-	-	0.12	0.60
S_12	S_3	-	-	0.15	0.60
S_13	S_3	-	-	0.18	0.59
S_14	S_3	-	-	0.16	0.59
S_15	S_3	-	-	0.07	0.59



(a) Vertical Approach

Figure 5.1 Association Rules for Recommendation within Top 20 in Confidence Ranking



(b) Horizontal Approach

Figure 5.1 Association Rules for Recommendation within Top 20 in Confidence Ranking

6 Management of Established Recommendation Rules over Time

In this section, we establish a scheme for managing derived association rules along the time axis, that is, some decision rules are given so as to determine whether one should continue to keep a derived association rule as valid over the next period, stop the rule but hold until the end of the next period for reexamination, or stop the rule and abandon.

More specifically, the conditions of $Supp(X,Y) \geq 0.1$ and $Conf(Y|X) \geq 0.1$ are adopted for judging whether or not $X \rightarrow Y$ is valid in any month. Based on such results for the last three months for $X \rightarrow Y$, the decision rules are given as in TABLE 6.1, concerning whether $X \rightarrow Y$ should be applied next month, should not be applied next month but be kept on hold until one month after the next month, or should be abandoned.

TABLE 6.1 Management Scheme for Association Rules

n-2	n-1	n	n+1
○	○	○	Apply
×	○	○	Apply
○	×	○	Hold
×	×	○	Hold
○	○	×	Hold
×	○	×	Hold
○	×	×	Abandon
×	×	×	Abandon

TABLE 6.2 below exhibits a dozen of derived association rules based on the vertical approach, where the values of Supp, Conf and Lift are shown for June, July and August. TABLE 6.3 shows a sequence of the decisions concerning the status of each of the association rules over the period of September through December, by applying the scheme defined in TABLE 6.1.

It should be noted that the dozen association rules in TABLE 6.2 involve 12 subcategories, where both $X \rightarrow Y$ and $Y \rightarrow X$ hold true for 6 different pairs (X,Y). For X= "Liquid Detergent" and Y = "Fabric Softener," one sees that $X \rightarrow Y$ and $Y \rightarrow X$ are quite comparable, while $X \rightarrow Y$ dominates $Y \rightarrow X$ for X= "Soap, Facial Cleansing Cream, Bath Additive" and Y = "Fabric Softener," or X = "Shampoo" and Y = "Fabric Softener."

When a derived association rule is decided to be applied in a month, we call it as a recommendation rule for that month. Only 4 recommendation rules, for example, continue to remain as a recommendation rule throughout the period from September to December, as can be seen from TABLE 6.3.

TABLE 6.2 Derivation of Association Rules

Rule		6			7			8		
X	*Y*	Supp	Conf	Lift	Supp	Conf	Lift	Supp	Conf	Lift
Liquid Detergent	Fabric Softener	0.365	0.678	1.191	0.385	0.694	1.172	0.392	0.702	1.161
Fabric Softener	Liquid Detergent	0.365	0.641	1.191	0.385	0.65	1.172	0.392	0.648	1.161
Soap, Facial Cleansing Cream, Bath Additive	Fabric Softener	0.277	0.604	1.062	0.298	0.627	1.059	0.319	0.639	1.056
Fabric Softener	Soap, Facial Cleansing Cream, Bath Additive	0.266	0.494	1.076	0.277	0.5	1.052	0.294	0.527	1.055
Shampoo	Fabric Softener	0.218	0.619	1.087	0	0	0	0.222	0.633	1.046
Fabric Softener	Shampoo	0.218	0.383	1.087	0	0	0	0.222	0.367	1.046
Bathtub Cleaner	Shampoo	0.106	0.379	1.078	0	0	0	0.11	0.378	1.078
Shampoo	Bathtub Cleaner	0.106	0.302	1.078	0	0	0	0.11	0.314	1.078
Repellents	Fabric Softener	0.105	0.516	0.906	0	0	0	0	0	0
Fabric Softener	Repellents	0.105	0.185	0.906	0	0	0	0	0	0
Tank Cleaner	Fabric Softener	0	0	0	0	0	0	0.123	0.203	1.044
Fabric Softener	Tank Cleaner	0	0	0	0	0	0	0.123	0.632	1.044

TABLE 6.3 Status Transition of Association Rules

Rule: $X \rightarrow Y$		6	7	8	9	10	11	12
X	Y							
Liquid Detergent	Fabric Softener	○	○	○	Apply	Apply	Apply	Apply
Fabric Softener	Liquid Detergent	○	○	○	Apply	Apply	Apply	Apply
Soap, Facial Cleansing Cream, Bath Additive	Fabric Softener	○	○	○	Apply	Apply	Apply	Apply
Fabric Softener	Soap, Facial Cleansing Cream, Bath Additive	○	○	○	Apply	Apply	Apply	Apply
Shampoo	Fabric Softener	○	×	○	Hold	Hold	Apply	Apply
Fabric Softener	Shampoo	○	×	○	Hold	Hold	Apply	Apply
Bathtub Cleaner	Shampoo	○	×	○	Hold	Hold	Apply	Apply
Shampoo	Bathtub Cleaner	○	×	○	Hold	Hold	Apply	Apply
Repellents	Fabric Softener	○	×	×	Abandon	Abandon	Abandon	Abandon
Fabric Softener	Repellents	○	×	×	Abandon	Abandon	Abandon	Abandon
Tank Cleaner	Fabric Softener	×	×	○	Hold	Hold	Apply	Hold
Fabric Softener	Tank Cleaner	×	×	○	Hold	Hold	Apply	Hold

7 Economic Impact of Recommendation Rules

Recommendation rules discussed in Section 6 are determined by exclusively focusing on frequencies of simultaneous purchasing across two different small categories over one month period, and the underlying economic aspect is completely ignored. In reality, however, the economic impact of a recommendation rule should play a centered role in deciding whether or not a sales campaign should take place along the recommendation rule. The purpose of this section is to establish an analytic framework for assessing the economic impact of a recommendation rule.

For two small categories X, Y, let $X \to Y$ be a recommendation rule for month $n+1$ derived through the procedure discussed in Section 6. In order to assess the economic impact of $X \to Y$, the following notation is employed.

$B_{X\to Y}(n)$: the average total purchasing amount of money per customer for the products in Y, where the average is taken over those customers who purchased products in both X and Y in month n.

$C_{X\to Y}(n)$: $Conf(Y|X)$ *for X→Y.*

$T_{\neg Y|X}(n)$: a set of customers who purchased at least one product in X but not in Y.

$N_{\neg Y|X}(\mathrm{n})$: the number of customers in $T_{\neg Y|X}(n)$.

$1-Q$: the discount rate offered to those customers in $T_{\neg Y|X}(n)$ through a sales campaign in month $n+1$.

$C_{X\to Y}(n)+\Delta(Q)$: the probability that a customer in $T_{\neg Y|X}(n)$ purchases a product in Y in month $n+1$ upon the sales campaign.

It should be noted that $C_{X\to Y}(n)$ represents the probability that a customer purchases at least one product in X in month n also purchases at least one product in Y in month n. Assuming that the underlying probabilistic structure is carried over to month $n+1$, $C_{X\to Y}(n)+\Delta(Q)$ describes the probability that a customer in $T_{\neg Y|X}(n)$ will buy a product in Y in month $n+1$.

For the incremental probability $\Delta(Q)$, it is natural to assume the following conditions.

① $C_{X\to Y}(n)+\Delta(Q)<1$
② $\Delta(Q)$: monotonically decreasing in Q
③ $\Delta(1)=0$ i.e. the incremental probability is zero if no discount is offered
④ $^{\exists}Q_0 : 0<Q_0<Q<1$: there exists the maximum discount $(1-Q_0)$ that can be offered because of the production cost

In this paper, we assume that

$$\Delta(Q)=\beta(1-Q)^{\alpha} \quad . \tag{2}$$

where $\alpha, \beta > 0$. When $\alpha=1$, $\Delta(Q)$ is a line. For $0<\alpha<1$, $\Delta(Q)$ is convex, where consumers are motivated to purchase even for a small discount but the discount effect diminishes as the discount rate increases. If $\alpha<1$, $\Delta(Q)$ is concave and consumers are affected only by large discounts. These three patterns are depicted in Figure 7.1.

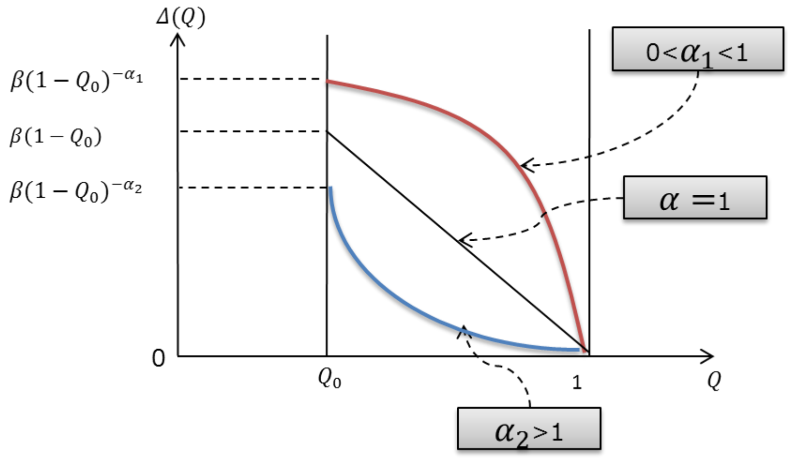


Figure 7.1 Three Patterns of $\Delta(Q)$

We are now in a position to discuss the economic impact of $X \rightarrow Y$. Given a discount rate $1-Q$ and a recommendation rule $X \rightarrow Y$ for month $n+1$, let $f_{X\rightarrow Y}(Q, n+1)$ be the economic impact of $X \rightarrow Y$ in month $n+1$. One then sees that

$$f_{X\rightarrow Y}(Q, n+1) = N_{\neg Y|X}(n) \times [C_{X\rightarrow Y}(n) + \Delta(Q)] \times B_{X\rightarrow Y}(n) \times Q \quad . \tag{3}$$

The discount rate $1-Q$ should be determined by solving the following optimization problem.

$$Q^* = argmax_{\ Q_0<Q<1} \{f_{X\rightarrow Y}(Q, n+1)\} \quad . \tag{4}$$

A sales campaign for the recommendation rule $X \rightarrow Y$ should then be implemented if and only if

$$Campaign\ Cost < f_{X\rightarrow Y}(Q^*, n+1) - f_{X\rightarrow Y}(0, n+1) \quad . \tag{5}$$

In order to demonstrate how the economic impact of a recommendation rule can be computed, we consider the first recommendation rule in TABLE 6.2 for the month of September. For $X =$ "Liquid Detergent" and $Y =$ "Fabric Softener" with $n = 9$ (September), it turns out that $N(\neg Y|X)(9) = 1{,}354$, $C_{X\rightarrow Y}(9) = 0.702$ and $B_{X\rightarrow Y}(n) = 962.39$. For $\Delta(Q)$, we assume that $\alpha = 0.1$ and $\beta = 0.05$ so that $\Delta(Q) = 0.1 \times\ (1-Q)^{0.05}$. Consequently, one has

$$f_{X\rightarrow Y}(Q, n+1) = 1{,}354 \times \left[0.702 + 0.1 \times\ (1-Q)^{0.05}\right] \times 962.39 \times Q \quad . \tag{6}$$

Suppose $Q_0 = 0.7$. Since $f_{X\rightarrow Y}(Q, n+1)$ in (6) is concave in Q, the global maximum point can be obtained by differentiating it with respect to Q and then setting the resulting formula to be 0, which yields 0.9523. Since $Q_0 = 0.7 < 0.9523$, the optimization problem in (3) can be solved at $Q^* = 0.9523$. This in turn enables one to evaluate the incremental probability as $\Delta(Q^*) = 0.086$. Finally, one can compute

$$f_{X\rightarrow Y}(Q^*, n+1) - f_{X\rightarrow Y}(0, n+1) = 977{,}844 - 914{,}759 = 63{,}085 \quad . \tag{7}$$

In conclusion, a sales campaign for $X \rightarrow Y$ should be implemented with the budget of $63{,}085$ or less.

8 Conclusion

In this paper, a framework was established for developing recommendation engines for DIY stores so as to enhance their sales. A set of real data for the study was first described, consisting of monthly sales records of 74 DIY stores across Japan. Through a basic analysis, a key DIY store of business importance was then selected for further study. In order to overcome the massive data volume, two sales areas were chosen, where the basis for the choice is the comparison of association rules with respect to simultaneous purchasing in two sales areas over the 6 month period. Recommendation rules were derived based on the conditions that both the support and the confidence are greater than 0.1. Also established was a management scheme concerning whether a recommendation rule should be applied again next month, should not be applied next month but kept on hold until one month after the next month, or should be abandoned. Finally, the problem of how to assess the economic impact of a recommendation rule was addressed.

The novelty of this paper can be summarized as follows.

- In developing recommendation engines, the vertical approach and the horizontal approach are proposed, where the former focuses on product categories while the latter is centered on a specific class of customers.
- A new scheme is designed for management of derived association rules concerning the decision of whether a recommendation rule should be applied again next month, should not be applied next month but kept on hold until one month after the next month, or should be abandoned.
- A new economic model is developed to assess the economic impact of a recommendation rule, thereby providing the budget boundary affordable for the necessary sales campaign.

The achievement of this paper enables DIY stores to device new sales campaign approaches. For example, upon realizing what specific items have been purchased by a specific customer, it is possible to print out certain coupons in the end of the receipt of the customer based on the judgment of the recommendation engine.

9 References

[1] Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl, "Analysis of recommendation algorithms for e-commerce," EC '00 Proceedings of the 2nd ACM conference on Electronic commerce, pp. 158-167.

[2] S.Wesley Changchien, Tzu-Chuen Lu,"Mining association rules procedure to support on-line recommendation by customers and products fragmentation," Expert Systems with Applications, Volume 20, Issue 4, May 2001, pp. 325–335.

[3] Zijian Zheng, Ron Kohavi, Llew Mason, "Real world performance of association rule algorithms," KDD '01 Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining,pp. 401-406.

[4] Weiyang Lin, Sergio A. Alvarez, Carolina Ruiz, "Efficient Adaptive-Support Association Rule Mining for Recommender Systems," Data Mining and Knowledge Discovery archive, Volume 6 Issue 1, January 2002, pp. 83-105.

[5] Choonho Kim, Juntae Kim, "A recommendation algorithm using multi-level association rules," Web Intelligence, 2003. WI 2003, pp. 524 – 527.

[6] Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan, Suh-Yin, "Emotion-based music recommendation by association discovery from film music," MULTIMEDIA '05 Proceedings of the 13th annual ACM international conference on Multimedia, pp. 507-510.

[7] Duen-Ren Liu, Ya-Yueh Shih, "Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences," Journal of Systems and Software archive, Volume 77 Issue 2, August 2005, pp 181-191.

[8] J. J. Sandvig, Bamshad Mobasher, Robin Burke, "Robustness of collaborative recommendation based on association rule mining," RecSys '07 Proceedings of the 2007 ACM conference on Recommender systems, pp. 105-112.

[9] Shyue-Liang Wang , Dipen Patel, Ayat Jafari, Tzung-Pei Hong, "Hiding collaborative recommendation association rules," Applied Intelligence, August 2007, Volume 27, Issue 1, pp. 67-77.