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].
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>
<thead>
<tr>
<th>TABLE 2.1 Definition of Store Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store Areas (1000㎡)</td>
</tr>
<tr>
<td>Small</td>
</tr>
<tr>
<td>Medium</td>
</tr>
<tr>
<td>Large</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 2.2 Compositions of Stores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Areas</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Small</td>
</tr>
<tr>
<td>Medium</td>
</tr>
<tr>
<td>Large</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>Gross Sales</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>Store 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20%</td>
<td>Store 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Store 3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30%</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>60%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>70%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>80%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>90%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>100%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Gross Sales (About ¥ 3 Billion)</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40%</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>60%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>70%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>80%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>100%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

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, let and be defined by

\[
N : \text{the number of customers who purchased at least once at Store 1 over the data period; and}
\]

\[
N(Z) : \text{the number of customers who purchased at least one of the products which belong to } Z \text{ at Store 1 over the data period.}
\]

For two different sales areas \(X\) and \(Y\), we also define \(\text{Supp}(X,Y)\) and \(\text{Conf}(X,Y)\) by

\[
\text{Supp}(X,Y) = \frac{N(X \cap Y)}{N}; \quad \text{Conf}(X,Y) = \frac{N(X \cap Y)}{N(X)}.
\]

It should be noted that, higher the value of \(\text{Supp}(X,Y)\) is, more likely that “simultaneous purchasing” exists between \(X\) and \(Y\). Furthermore, if \(\text{Conf}(X,Y) > \text{Conf}(Y,X)\), 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 \(\text{Supp}\) and \(\text{Conf}\). The thickness of the arrow from \(X\) to \(Y\) indicates the strength of \(\text{Supp}(X,Y)\), which is written as \(S\) in the middle of the three numbers on the arrow. The first number \(C_1\) is the value of \(\text{Conf}(Y|X)\) and the last number \(C_2\) is that of \(\text{Conf}(Y|X)\) with \(C_1 > C_2\).

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 (\(S = 0.54 \text{ and } C_1 = 0.9\)).
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.

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 \textit{HD-SAG} (Sales Area Group) be the set of sales areas that belong to Household Goods and Daily Necessities. We define \textit{HD-MCG} (Medium Category Group) similarly for medium categories. The purpose of this section is first to select medium categories in \textit{HD-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-SAG*. One sees that 12 medium categories out of 37, or 32% of the total medium categories in *HD-SAG*, collectively account for 80% of both the gross sales and the number of products purchased for *HD-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*

<table>
<thead>
<tr>
<th>Gross Sales</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20%</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30%</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>40%</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>50%</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>60%</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>70%</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>9</td>
<td>11</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>80%</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>90%</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>100%</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>19</td>
<td>37</td>
</tr>
</tbody>
</table>

---

**Figure 3.1** Interrelations among 12 Selected Medium Categories
(Threshold: Supp ≥ 0.25; Conf ≥ 0.1)
TABLE 3.2     Selected Medium Categories in HD-AG

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

Let $HD-MC-SEL$ 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 $HD-MC-SEL$. 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 $HD-MC-SEL$. The resulting set of customers is denoted by $HD-MC-SEL-C-80$. We now focus on $HD-MC-SEL$ and $HD-MC-SEL-C-80$ 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

<table>
<thead>
<tr>
<th>HD-MC-SEL</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>20%</td>
<td>99</td>
<td>277</td>
<td>307</td>
<td>314</td>
<td>321</td>
<td>324</td>
<td>324</td>
<td>324</td>
<td>324</td>
<td>324</td>
</tr>
<tr>
<td>30%</td>
<td>115</td>
<td>569</td>
<td>880</td>
<td>991</td>
<td>1035</td>
<td>1054</td>
<td>1060</td>
<td>1064</td>
<td>1066</td>
<td>1067</td>
</tr>
<tr>
<td>40%</td>
<td>117</td>
<td>679</td>
<td>1318</td>
<td>1731</td>
<td>1950</td>
<td>2025</td>
<td>2060</td>
<td>2080</td>
<td>2088</td>
<td>2093</td>
</tr>
<tr>
<td>50%</td>
<td>118</td>
<td>706</td>
<td>1513</td>
<td>2301</td>
<td>2860</td>
<td>3169</td>
<td>3310</td>
<td>3380</td>
<td>3407</td>
<td>3417</td>
</tr>
<tr>
<td>60%</td>
<td>119</td>
<td>715</td>
<td>1566</td>
<td>2590</td>
<td>3545</td>
<td>4303</td>
<td>4753</td>
<td>4966</td>
<td>5052</td>
<td>5096</td>
</tr>
<tr>
<td>70%</td>
<td>119</td>
<td>722</td>
<td>1590</td>
<td>2703</td>
<td>3943</td>
<td>5236</td>
<td>6217</td>
<td>6872</td>
<td>7146</td>
<td>7250</td>
</tr>
<tr>
<td>80%</td>
<td>120</td>
<td>727</td>
<td>1600</td>
<td>2734</td>
<td>4115</td>
<td>5790</td>
<td>7446</td>
<td>8947</td>
<td>9815</td>
<td>10120</td>
</tr>
<tr>
<td>90%</td>
<td>121</td>
<td>732</td>
<td>1610</td>
<td>2757</td>
<td>4177</td>
<td>5964</td>
<td>8153</td>
<td>1081</td>
<td>13199</td>
<td>14333</td>
</tr>
<tr>
<td>100%</td>
<td>122</td>
<td>736</td>
<td>1626</td>
<td>2780</td>
<td>4238</td>
<td>6064</td>
<td>8398</td>
<td>11515</td>
<td>16066</td>
<td>39477</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Number of Products Purchased</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20%</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>30%</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>50%</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>60%</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>70%</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>14</td>
<td>14</td>
<td>17</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>14</td>
<td>17</td>
<td>21</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>90%</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>15</td>
<td>20</td>
<td>26</td>
<td>34</td>
<td>38</td>
</tr>
<tr>
<td>100%</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>15</td>
<td>20</td>
<td>27</td>
<td>36</td>
<td>69</td>
</tr>
</tbody>
</table>

TABLE 3.5  Selected Small Categories within $HD\text{-}MC\text{-}SEL$

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

Copyright © by IAI. Unauthorized reproduction of this article is prohibited.
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-SC-SEL in TABLE 3.5 with focus on the customer class HD-SC-SEL-C-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 $X \rightarrow Y$ with $Supp \geq 0.01; Conf \geq 0.01$

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

<table>
<thead>
<tr>
<th>HD-HORI-C</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>20%</td>
<td>107</td>
<td>334</td>
<td>379</td>
<td>396</td>
<td>408</td>
<td>418</td>
<td>422</td>
<td>424</td>
<td>424</td>
<td>429</td>
</tr>
<tr>
<td>30%</td>
<td>118</td>
<td>619</td>
<td>960</td>
<td>1095</td>
<td>1156</td>
<td>1191</td>
<td>1216</td>
<td>1230</td>
<td>1238</td>
<td>1247</td>
</tr>
<tr>
<td>40%</td>
<td>119</td>
<td>705</td>
<td>1397</td>
<td>1864</td>
<td>2115</td>
<td>2211</td>
<td>2275</td>
<td>2318</td>
<td>2342</td>
<td>2364</td>
</tr>
<tr>
<td>50%</td>
<td>120</td>
<td>716</td>
<td>1562</td>
<td>2434</td>
<td>3072</td>
<td>3372</td>
<td>3571</td>
<td>3691</td>
<td>3762</td>
<td>3810</td>
</tr>
<tr>
<td>60%</td>
<td>121</td>
<td>722</td>
<td>1596</td>
<td>2679</td>
<td>3777</td>
<td>4624</td>
<td>5114</td>
<td>5409</td>
<td>5568</td>
<td>5655</td>
</tr>
<tr>
<td>70%</td>
<td>121</td>
<td>727</td>
<td>1607</td>
<td>2741</td>
<td>4091</td>
<td>5583</td>
<td>6674</td>
<td>7425</td>
<td>7801</td>
<td>8006</td>
</tr>
<tr>
<td>80%</td>
<td>122</td>
<td>730</td>
<td>1614</td>
<td>2799</td>
<td>4212</td>
<td>5966</td>
<td>7910</td>
<td>9640</td>
<td>10645</td>
<td>11126</td>
</tr>
<tr>
<td>90%</td>
<td>122</td>
<td>735</td>
<td>1619</td>
<td>2771</td>
<td>4228</td>
<td>6044</td>
<td>8360</td>
<td>11280</td>
<td>14261</td>
<td>15689</td>
</tr>
<tr>
<td>100%</td>
<td>122</td>
<td>736</td>
<td>1626</td>
<td>2780</td>
<td>4238</td>
<td>8398</td>
<td>11515</td>
<td>16066</td>
<td>39477</td>
<td></td>
</tr>
</tbody>
</table>

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 $\text{Supp}(X,Y) \geq 0.01$ and $\text{Conf}(X,Y) \geq 0.01$ again. Some of the resulting recommendation rules are listed in TABLE 4.3 in the order of $\text{Conf}(X,Y)$.

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

<table>
<thead>
<tr>
<th>HD-HORI-S</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>14</td>
<td>17</td>
<td>17</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>30%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>16</td>
<td>18</td>
<td>21</td>
<td>27</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>17</td>
<td>22</td>
<td>26</td>
<td>37</td>
<td>40</td>
<td>41</td>
</tr>
<tr>
<td>50%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>18</td>
<td>25</td>
<td>33</td>
<td>46</td>
<td>56</td>
<td>58</td>
</tr>
<tr>
<td>60%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>18</td>
<td>25</td>
<td>34</td>
<td>48</td>
<td>67</td>
<td>85</td>
</tr>
<tr>
<td>70%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>18</td>
<td>25</td>
<td>34</td>
<td>48</td>
<td>73</td>
<td>187</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>18</td>
<td>25</td>
<td>34</td>
<td>48</td>
<td>73</td>
<td>187</td>
</tr>
<tr>
<td>90%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>18</td>
<td>25</td>
<td>34</td>
<td>48</td>
<td>73</td>
<td>187</td>
</tr>
<tr>
<td>100%</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>18</td>
<td>25</td>
<td>34</td>
<td>48</td>
<td>73</td>
<td>187</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>ID</th>
<th>Small Category Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>Bleach Cleaner</td>
</tr>
<tr>
<td>S_2</td>
<td>Liquid Detergent</td>
</tr>
<tr>
<td>S_3</td>
<td>Fabric Softener</td>
</tr>
<tr>
<td>S_4</td>
<td>Bath Tub Cleaner</td>
</tr>
<tr>
<td>S_5</td>
<td>Dishwashing Detergent</td>
</tr>
<tr>
<td>S_6</td>
<td>Laundry Aids</td>
</tr>
<tr>
<td>S_7</td>
<td>Shampoo</td>
</tr>
<tr>
<td>S_8</td>
<td>Toothpaste</td>
</tr>
<tr>
<td>S_9</td>
<td>Soap, Facial Cleansing Cream, Bath Additive</td>
</tr>
<tr>
<td>S_10</td>
<td>Bath Additive</td>
</tr>
<tr>
<td>S_11</td>
<td>Toilet Cleaner</td>
</tr>
<tr>
<td>S_12</td>
<td>Wet Tissue</td>
</tr>
<tr>
<td>S_13</td>
<td>Cling Film and Foil</td>
</tr>
<tr>
<td>S_14</td>
<td>Recycled Toilet Paper</td>
</tr>
<tr>
<td>S_15</td>
<td>Powder Detergent</td>
</tr>
<tr>
<td>S_16</td>
<td>Wool Washing Detergent</td>
</tr>
<tr>
<td>S_17</td>
<td>Children Toothbrush</td>
</tr>
<tr>
<td>S_18</td>
<td>Sanitary Items for Night Use</td>
</tr>
<tr>
<td>S_19</td>
<td>Sanitary Items for Day Use</td>
</tr>
<tr>
<td>S_20</td>
<td>Facial Cleansing Cream</td>
</tr>
<tr>
<td>S_21</td>
<td>Rinse Treatments</td>
</tr>
<tr>
<td>S_22</td>
<td>Other Sanitary Items</td>
</tr>
<tr>
<td>S_23</td>
<td>Diaper</td>
</tr>
<tr>
<td>S_24</td>
<td>Garbage Bag</td>
</tr>
</tbody>
</table>

We note from TABLE 4.3 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 \( \text{Supp} \geq 0.01 \) and \( \text{Conf} \geq 0.01 \). 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  (b) Horizontal Approach

<table>
<thead>
<tr>
<th>Number of Rules</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>Number of Rules</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical Approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Horizontal Approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>S_3</td>
<td></td>
<td></td>
<td>10%</td>
<td>S_2</td>
<td>S_3</td>
<td></td>
<td>S_7</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20%</td>
<td>S_3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>S_2</td>
<td></td>
<td></td>
<td>30%</td>
<td>S_9</td>
<td></td>
<td>S_23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>S_9</td>
<td>S_7</td>
<td></td>
<td>40%</td>
<td></td>
<td>S_24</td>
<td>S_5</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical</td>
<td>392</td>
<td>416</td>
<td>377</td>
<td>482</td>
<td>377</td>
</tr>
<tr>
<td>Horizontal</td>
<td>2225</td>
<td>2152</td>
<td>2091</td>
<td>1732</td>
<td>2038</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Supp</th>
<th>Conf</th>
<th>Supp</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>S_2</td>
<td>0.14</td>
<td>0.63</td>
<td>0.17</td>
<td>0.66</td>
</tr>
<tr>
<td>S_1</td>
<td>S_3</td>
<td>0.15</td>
<td>0.69</td>
<td>0.19</td>
<td>0.72</td>
</tr>
<tr>
<td>S_2</td>
<td>S_3</td>
<td>0.29</td>
<td>0.65</td>
<td>0.36</td>
<td>0.68</td>
</tr>
<tr>
<td>S_7</td>
<td>S_3</td>
<td>0.17</td>
<td>0.59</td>
<td>0.22</td>
<td>0.62</td>
</tr>
<tr>
<td>S_10</td>
<td>S_3</td>
<td>0.06</td>
<td>0.58</td>
<td>0.08</td>
<td>0.6</td>
</tr>
<tr>
<td>S_6</td>
<td>S_3</td>
<td>0.08</td>
<td>0.58</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_16</td>
<td>S_3</td>
<td>0.04</td>
<td>0.67</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_17</td>
<td>S_3</td>
<td>0.02</td>
<td>0.64</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_17</td>
<td>S_9</td>
<td>0.02</td>
<td>0.60</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_17</td>
<td>S_2</td>
<td>0.02</td>
<td>0.61</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_18</td>
<td>S_3</td>
<td>0.03</td>
<td>0.63</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_19</td>
<td>S_3</td>
<td>0.06</td>
<td>0.62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_21</td>
<td>S_3</td>
<td>0.04</td>
<td>0.60</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_22</td>
<td>S_3</td>
<td>0.03</td>
<td>0.58</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_22</td>
<td>S_2</td>
<td>0.03</td>
<td>0.58</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S_4</td>
<td>S_3</td>
<td>-</td>
<td>-</td>
<td>0.17</td>
<td>0.62</td>
</tr>
<tr>
<td>S_5</td>
<td>S_3</td>
<td>-</td>
<td>-</td>
<td>0.23</td>
<td>0.62</td>
</tr>
<tr>
<td>S_6</td>
<td>S_2</td>
<td>-</td>
<td>-</td>
<td>0.09</td>
<td>0.60</td>
</tr>
<tr>
<td>S_8</td>
<td>S_3</td>
<td>-</td>
<td>-</td>
<td>0.13</td>
<td>0.61</td>
</tr>
<tr>
<td>S_8</td>
<td>S_9</td>
<td>-</td>
<td>-</td>
<td>0.13</td>
<td>0.58</td>
</tr>
<tr>
<td>S_11</td>
<td>S_3</td>
<td>-</td>
<td>-</td>
<td>0.12</td>
<td>0.60</td>
</tr>
<tr>
<td>S_12</td>
<td>S_3</td>
<td>-</td>
<td>-</td>
<td>0.15</td>
<td>0.60</td>
</tr>
<tr>
<td>S_13</td>
<td>S_3</td>
<td>-</td>
<td>-</td>
<td>0.18</td>
<td>0.59</td>
</tr>
<tr>
<td>S_14</td>
<td>S_3</td>
<td>-</td>
<td>-</td>
<td>0.16</td>
<td>0.59</td>
</tr>
<tr>
<td>S_15</td>
<td>S_3</td>
<td>-</td>
<td>-</td>
<td>0.07</td>
<td>0.59</td>
</tr>
</tbody>
</table>
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 $\text{Supp}(X, Y) \geq 0.1$ and $\text{Conf}(Y|X) \geq 0.1$ 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

<table>
<thead>
<tr>
<th>n-2</th>
<th>n-1</th>
<th>n</th>
<th>n+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>○</td>
<td>○</td>
<td>○</td>
<td>Apply</td>
</tr>
<tr>
<td>×</td>
<td>○</td>
<td>○</td>
<td>Apply</td>
</tr>
<tr>
<td>○</td>
<td>×</td>
<td>○</td>
<td>Hold</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>○</td>
<td>Hold</td>
</tr>
<tr>
<td>○</td>
<td>○</td>
<td>×</td>
<td>Hold</td>
</tr>
<tr>
<td>×</td>
<td>○</td>
<td>×</td>
<td>Hold</td>
</tr>
<tr>
<td>○</td>
<td>×</td>
<td>×</td>
<td>Abandon</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>Abandon</td>
</tr>
</tbody>
</table>

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 $X = \text{"Liquid Detergent"}$ and $Y = \text{"Fabric Softener,"}$ one sees that $X \rightarrow Y$ and $Y \rightarrow X$ are quite comparable, while $X \rightarrow Y$ dominates $Y \rightarrow X$ for $X = \text{"Soap, Facial Cleansing Cream, Bath Additive"}$ and $Y = \text{"Fabric Softener,"}$ or $X = \text{"Shampoo"}$ and $Y = \text{"Fabric Softener."}$

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.
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
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 \( X \rightarrow Y \) be a recommendation rule for month \( n + 1 \) derived through the procedure discussed in Section 6. In order to assess the economic impact of \( X \rightarrow Y \), the following notation is employed.

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

\[
C_{X \rightarrow Y}(n) : \text{Conf}(Y|X) \text{ for } X \rightarrow Y.
\]

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

\[
N_{n \setminus Y|X}(n) : \text{the number of customers in } T_{n \setminus Y|X}(n).
\]

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

\[
C_{X \rightarrow Y}(n) + \Delta(Q) : \text{the probability that a customer in } T_{n \setminus Y|X}(n) \text{ purchases a product in } Y \text{ in month } n + 1 \text{ upon the sales campaign.}
\]

It should be noted that \( C_{X \rightarrow Y}(n) \) 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 \( n + 1 \), \( C_{X \rightarrow Y}(n) + \Delta(Q) \) describes the probability that a customer in \( T_{n \setminus Y|X}(n) \) 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. \( C_{X \rightarrow Y}(n) + \Delta(Q) < 1 \)
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

\[
\Delta(Q) = \beta(1 - Q)^\alpha.
\]

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.
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

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

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

$$Q^* = \operatorname{argmax}_{0 \leq Q < 1} \{ f_{X \rightarrow Y}(Q, n + 1) \} \ .$$  \hspace{1cm} (4)

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

$$\textit{Campaign Cost} < f_{X \rightarrow Y}(Q^*, n + 1) - f_{X \rightarrow Y}(0, n + 1) \ .$$  \hspace{1cm} (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 $N_{=Y|X}(9) = 1,354 \ , \ C_{X \rightarrow Y}(9) = 0.702$ 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

$$f_{X \rightarrow Y}(Q, n + 1) = 1,354 \times [0.702 + 0.1 \times (1 - Q)^{0.05}] \times 962.39 \times Q \ .$$  \hspace{1cm} (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

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


