Application of Sequence Alignment Method to Product Assortment and Shelf Space Allocation

Peiqian Liu, Hairu Guo, Weipeng An
School of Computer Science and Technology/Henan Polytechnic University, Jiaozuo, China
liupeiqian@hpu.edu.cn

Abstract—In retailing, decisions about product assortment and shelf space allocation have a significant effect on customers’ purchasing decisions. Traditionally, researchers usually employed the space elasticity to optimize product assortment and space allocation models. However, the large number of parameters requiring to estimate and the non-linear nature of space elasticity can reduce the efficacy of the space elasticity based models. Instead of space elasticity, this paper utilizes a data mining approach, Sequence Alignment Method (SAM), to resolve the product assortment and allocation problems in retailing. In our approach, the SAM is applied to explore the relationships between product categories and is compared to association rule mining. Experimental results show that SAM achieves better quality than that of association rule mining and can generate very useful information to shelf space management.

Index Terms—Data mining; Shelf space management; SAM; Product taxonomy

I. INTRODUCTION

Shelf space is an important resource for retail stores since a great quantity of products compete the limited shelf space for display. Retailers need frequently make decisions about which products to display (assortment) and how much shelf space to allocate these products (allocation) [1]. Product assortment and shelf space allocation are two important issues in retailing which can affect the customers’ purchasing decisions. Through the proficient shelf space management, retailers can improve return on inventory and consumer’s satisfaction, and therefore increase sales and margin profit [2].

Traditionally, researchers apply the space elasticities to determine which products to stock and how much shelf space to allocate in each product category. However, there are two major limitations that reduce the effectiveness of the space elasticity [3]. First, due to the non-linear nature of space elasticity, the space elasticity-based models are very complicated, and the specific solution approach is developed for each model. Additionally, it is necessary to estimate a large number of parameters by using the space elasticity.

With the rapid development of information technology, transaction data can be easily collected through the point of sale (POS) system. The relationships between products hidden in transaction data can be discovered through data mining to assist product assortment and shelf space allocation [4]. It is not necessary to conduct a series of experiments to estimate a great quantity of parameters in space elasticities.

In this paper, we propose a data mining approach, called Sequence Alignment Method (SAM), to make decisions about which products to stock, how much shelf space allocated to the stocked products and where to display them. SAM analyzes customers’ shopping behaviors from transaction data to obtain relationships between product categories. As a dimensionality reduction technique, we employ a product taxonomy. In this taxonomy, similar products are identified and grouped together using the product taxonomy so as to build the customer profiles and to search for the neighbors in the reduced dimensional space. The products and categories frequently bought together can be displayed together. Finally, experiment shows SAM is better quality than association rule mining.

II. METHODOLOGY

The proposed procedure of shelf space management is divided into three phases: product classification, neighborhood formation, and top-N products generation.

A. Product classification

In this phase, the marketing manager or domain expert categorizes all the products by specifying the level of product aggregation on the product taxonomy. A product taxonomy is practically represented as a tree that classifies a set of products at a low level into a more general product at a higher level. The leaves of the tree denote the product instances, Stock Keeping Units (SKUs) in retail jargon, and non-leaf nodes denote product classes obtained by combining several nodes at a lower level into one parent node. The root node labeled by All denotes the most general product class. Fig.1 shows an example of product classification, where class nodes are denoted by shaded regions.

In this example, class(SKU00) = Outerwear and class(SKU10) = Shoes, etc.

Recent data mining research has shown that data mining algorithms usually produce the best results when product-related transactions are evenly occurred [5].
B. neighborhood formation

This phase performs computing the similarity between customers and, based on that, forming a neighborhood between a target customer and a number of like-minded customers.

A sequence is a number of elements arranged or coming one after the other in succession. In general, the distance (or similarity) between sequences is reflected by the number of operations necessary to convert one sequence into the other. As a result, SAM distance measure is represented by a score. The higher/lower the score, the more/less effort it takes to equalize the sequences and the less/more similar sequences are. In addition, SAM scores for the Insertion and the deletion operations to unique elements of source (first) and target (second) sequences during equalization process, not scores the reordering operations to common elements. Common elements appear in both compared sequences whereas unique elements appear in either one of them. Finally, SAM represents the minimum cost for equalizing two sequences.

In this paper, SAM distance measure between two sequences \( S_1 \) and \( S_2 \) is calculated using the following formula:

\[
d_{\text{SAM}}(S_1, S_2) = \sum_{i=1}^{n} w_i |A_{i1} - A_{i2}|
\]  

(1)

where \( d_{\text{SAM}} \) is the distance between two sequences \( S_1 \) and \( S_2 \); \( n \) is the length of \( S_1 \) or \( S_2 \) after equalization; \( w_i \) the weight value for the deletion operations or insertion operations on the \( i \)th element, a positive constant not equal to 0, determined by the researcher (\( w_i > 0 \)); \( A_{ij} \) is the shopping sum on the \( i \)th element in \( S_j \) (\( i = 1, 2 \)).

To illustrate SAM, consider the following sequences \( S_1 \) (source sequence) and \( S_2 \) (target sequence). Both sequences represent a purchased list extracted from the transaction database. Each element in the list is represented by a pair of number \( (C_i, A_i) \), where \( C_i \) represents the class ID, and \( A_i \) the shopping sum for \( C_i \).

\( S_1: \{(1,4), (2,0), (3,0), (4,1), (7,9), (8,9)\} \)

\( S_2: \{(1,1), (2,4), (3,1), (4,9), (7,4)\} \)

First, the maximum number of similar elements having the same class ID is defined. Then, in order to equalize \( S_1 \) with \( S_2 \); unique elements (2,0) and (3,0) are inserted into \( S_1 \) which gives the following sequences (a 0 means class 2 or 3 are not purchased):

\( S_1: \{(1,4), (2,0), (3,0), (4,1), (7,9), (8,9)\} \)

\( S_2: \{(1,1), (2,4), (3,1), (4,9), (7,4)\} \)

The equalization process continues with reordering common class 7 (or 8) in \( S_1 \) or \( S_2 \):

\( S_1: \{(1,4), (2,0), (3,0), (4,1), (7,9), (8,9)\} \)

\( S_2: \{(1,1), (2,4), (3,1), (4,9), (7,4), (8,9)\} \)

Finally, equalizing \( S_1 \) with \( S_2 \); took 2 insertion, (If we assign 2 to \( w_i \) for insertion and deletion, and 1 to \( w_i \) for otherwise) which gives us:

\[
d_{\text{SAM}}(S_1, S_2) = (4-1)+2 \times (4-0)+ 2 \times (1-0)+(9-1) + (9-4) + (9-4) = 26.
\]

After pair wise distances between sequences are calculated using SAM, A distance matrix is build for holding distance scores between each sequence pair on the diagonal. Because this study is focused on SAM, no special attention is paid to the clustering method. Therefore, a simple hierarchical clustering algorithm like Ward [6] is used to form neighborhood. In order to define an optimal solution for the number of neighborhood, \( r^2 \) index is used. \( r^2 \) is one of the most commonly used stop criteria [7], equals the proportion of variation and ranges in value from 0 to 1.

C. Top-N products generation

The final phase is to ultimately derive the top-N products from the neighborhood of customers. For each neighborhood, we produce a list of N products that the neighborhood is most likely to purchase. In this paper, we adopt the highest likely-to-buy rate (HLR) for generating a product list for a given neighborhood. The HLR method chooses products with the highest likely-to-buy rate of all neighbors. Formally, the likely-to-buy rate of the neighborhood \( i \) for a product \( j \) \( LTB_{i,j} \), is defined below:

\[
LTB_{i,j} = \frac{\sum_{j\in \text{neighborhood}(i)} r_{ij}}{\sum_{i\in \text{neighborhood}(i)} r_{ij}}
\]

(2)

Where \( r_{ij} \) is the total number of occurrences of purchases of a neighborhood \( i \) for a product \( j \); and \( r_i \) the total number of occurrences of all purchased of a neighborhood \( i \).

III. EXPERIMENTAL EVALUATION

A. Experimental result

The proposed data mining based procedure for product assortment and allocation is implemented with an example of a retail store. The database includes product data, customer data and transaction records. There are 3060 product items, which are divided into 32 categories. 10 percent of the transaction records was set as training data and 90 percent was set as test data. For neighborhood formation, \( r^2 \) index is used based on SAM distance matrix. In our research, 6 neighborhoods are formed with \( r^2 \) equal to 0.64.

Finally, top-N product list are generated for each neighborhood:

\[
N_1 = \{C_{11}, C_{12}, C_{16}, C_{18}, C_{20}, C_{27}, C_{30}\}
\]

\[
N_2 = \{C_{2}, C_{4}, C_{9}, C_{14}, C_{15}, C_{28}, C_{30}\}
\]

\[
N_3 = \{C_{1}, C_{4}, C_{8}, C_{19}, C_{21}, C_{29}, C_{30}\}
\]

\[
N_4 = \{C_{6}, C_{12}, C_{17}, C_{22}, C_{23}, C_{25}, C_{1}\}
\]

\[
N_5 = \{C_{3}, C_{7}, C_{10}, C_{11}, C_{18}, C_{32}, C_{21}\}
\]

\[
N_6 = \{C_{13}, C_{15}, C_{22}, C_{25}, C_{27}, C_{31}\}
\]

Figure 1. Example of product classification.
Where $N_i$ denotes neighborhood $i$, and $C_i$ represents the class ID of products. So classes in the same $N_i$ should be displayed as near as possible in a store.

**B. Evaluation metrics**

To evaluate the quality of the method, recall and precision have been widely used in relative research. Another widely used combination metric called F1 metric [8] that gives equal weight to both recall and precision was employed for our evaluation. It is computed as follows:

$$F1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$  

Finally, we compared the quality of SAM with that of association rule mining. Fig.2 shows our experimental results. It can be observed from Fig.2 that SAM works better than the association rule mining, achieving an average improvement of 38%.

![Quality comparison of SAM and association rule mining](image)

**IV. CONCLUSION**

With the rapid development of information technology, retailers have put a huge amount of transaction data in storage, and they potentially can be used to support shelf space management. This paper develops a data mining based approach SAM to simultaneously make decisions about Product Assortment and shelf space allocation. Firstly, the marketing manager categorizes all the products on the product taxonomy. Secondly, SAM is used to compute the similarity between customers and, based on that, forming a neighborhood between a number of like-minded customers. Finally, top-N product list are generated for each neighborhood. The top-N products can be displayed as near as possible in a store. Experimental results shows that SAM works better than the association rule mining.

**REFERENCES**


