Grocery Customer Behavior Analysis using RFID-based Shopping Paths Data

In-Chul Jung, Young S. Kwon

***Abstract*—**Knowing about the customer behavior in a grocery has been a long-standing issue in the retailing industry. The advent of RFID has made it easier to collect moving data for an individual shopper's behavior. Most of the previous studies used the traditional statistical clustering technique to find the major characteristics of customer behavior, especially shopping path. However, in using the clustering technique, due to various spatial constraints in the store, standard clustering methods are not feasible because moving data such as the shopping path should be adjusted in advance of the analysis, which is time-consuming and causes data distortion. To alleviate this problem, we propose a new approach to spatial pattern clustering based on the longest common subsequence. Experimental results using real data obtained from a grocery confirm the good performance of the proposed method in finding the hot spot, dead spot and major path patterns of customer movements.

***Keywords*—**customer path, shopping behavior, exploratory analysis, LCS, RFID

I. Introduction

THE goal of retailers (discount stores, department stores, convenience stores, supermarkets, etc) is to increase the gross profit margin through sales and cost reduction. This requires improving the efficiency of operation and providing attractive services for customers. Especially, the market focus of large discount stores has been continuous low price sales in tandem with the expansion of new branch stores. Recently, however, they have struggled with the decreased consumer spending due to the economic recession. This has removed the competitive position of a low price strategy only. This situation necessitates new marketing strategies such as aggressive promotions to customers. Traditional strategies include basket analysis or regional analysis based on customer purchase history and demographic information. Information about interested products is analyzed from customer purchase history and products recommended to customers through customer segmentation. The location of future profitable stores is identified using demographic information and regional analysis. However, more aggressive promotional activities are needed as these traditional analyses do not provide sufficient information to understand customer shopping patterns and behaviors in the physical store environment.

According to the research of Newman et al. [13] on customers' consumption decision making processes, the store environment (in-store layouts, product placement, etc.) closely affects customer consumption behavior. Therefore, improving the in-store environment is an important consequence of increased understanding of customer behavior. If we can determine the store areas where most sales activities occur and where customers tend to stay for a long time, we can decide where to display products and how to build an effective store environment. A more effective store environment can provide convenient services for customers and hence increases sales. Up to now, however, store managers have relied on experiences of the high-sales locations and those where customers tend to stay for a long time. Based on their experience, they decided where to display products and how to change in-store layout.

Farley and Ring [3] recorded the movement of some customers by following them in order to analyze the shopping path, one of the shopping behaviors, of customers. However, due to the numerous customers that visit the stores daily, it is difficult to record individual consumption behavior with only a few researchers and limited budget. The record is also not reliable due to the small sample number. Several researchers [4][9][11] have tried to solve these problems by using radio-frequency identification (RFID) and clustering techniques to analyze customers' shopping behavior, especially shopping path. However, the clustering method in customer shopping path suffers from two problems. First, during the process of clustering, the clusters which are divided at the location of obstacles such as sales shelves can converge into the same cluster group. Because people cannot cross obstacles such as shelves and merchandise stands, the store's physical environment and obstructions (shelves, merchandise stands, etc.) should be considered as a constraint for the shopping path clustering. Second, the length of a shopping path must be constant in order to apply a clustering algorithm to the shopping path data; however, this length is actually variable in a store.

Insufficient consideration has been given to the physical obstacles such as sales shelves and product displays. Furthermore, the input lengths between objects have to be identical to apply cluster techniques, but customers' traveling distances in stores vary. For example, customer "A" may finish shopping in 5 minutes and leave the store, but customer "B" may roam around the store for more than 30 minutes. Due to this wide variety in shopping courses between customers, we need to generalize the paths into a normalized shopping length based on time or space in order to apply the clustering algorithm. However, this normalizing process can introduce a distortion or noise into the shopping path or main travel information.

In-Chul Jung is with Department of Industrial & Systems Engineering, Dongguk University, Seoul, Korea (uhhaha@gmail.com).

Young S. Kwon is with Department of Industrial & Systems Engineering, Dongguk University, Seoul, Korea (yskwon@dgu.edu).

Therefore, in this paper we propose a new method of spatial patterns clustering in order to solve the problems of previous studies by changing the real shopping path to path location sequences and by using a new similarity measure between different customers' shopping paths. By adopting the longest common subsequence (LCS) method as the basic idea and expanding on it, we developed the main shopping path identification technique that is capable of identifying the hotspots where most of the customers' visits are made and the dead spots with few visits. We finally applied our newly developed method to the real data of a large supermarket store located in Seoul and analyzed its customer flow information.

II. Related Literature

The retailers of large supermarkets have profited by supplying a massive number of products at low price, and have recently begun to use various marketing strategies to gain a competitive advantage over other stores. They have tried to understand customers by analyzing demographic and transaction data. However, in order to understand customer shopping patterns inside a store, we need more data about customer behaviors. Transaction data and personal information only provide basic information such as how many times and when they have visited due to the limitation on customer behavior data collection technique.

Some researchers have studied cases of customer behaviors using direct observation or questionnaires. Harris [5] and McClure and West [12] have investigated whether store display and product brand change affect sales figures. Cox [1] studied some factors on unplanned purchases. Dickson and Sawyer [2] studied information processing and decision making in a store environment, and Hoyer [8] studied whether time pressure affected unplanned purchases and changes to brands and products. However, these studies did not investigate all customers but merely a sample as complete sampling would be very difficult and expensive. In recent years, technological advances such as inexpensive RFID and video cameras have been applied to the analysis of customer behavior [7][14][11]. These studies explained the behavioral properties of customers in the store and supported the formulation of marketing strategy and optimal operation. In particular Larsn et al. [11] and Hui et al. [10] tried to identify the major shopping path using RFID technology and the K-medoids clustering technique using the Euclidean distance. Existing research mainly used clustering techniques among data mining techniques to detect the main shopping path patterns. However, the use of clustering techniques based on Euclidean distance can reduce its accuracy due to obstacles such as sales stands and shelves in the actual store environment. For example, as shown in Fig. 1, if we measure the Euclidean distance from position a to b and c, position c is closer than position b. However, the actual customer's travel distance in the store from position "a" to position "c" is further than position "a" to position "b" because a customer has to walk around the sales stands. This is one of the reasons why Euclidean distance measurement in the existing clustering method is inadequate for in-store shopping path pattern grouping.

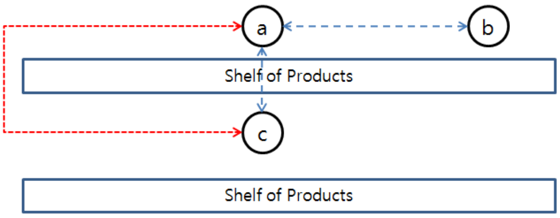

Fig. 1 Euclidean distance and actual customer path in a store

Since each customer has a different travel distance for shopping, we need to convert the different distances into one identical number throughout the entire process in order to apply them into a clustering algorithm. For example, when a clustering method like K-mediods is used, the input values have to be identical because of the characteristic of the algorithm. Therefore, different customers' different distances need to be converted into an identical number, and the normalization is applied to equalize the size of the trace. As in Fig. 2, we can choose between temporal normalization and spatial normalization. During the process, since a different value from the original distance is used as the input, the original information can be either deformed or lost. Therefore, this research provides a technique for detecting a main customer shopping path pattern in which all these path characteristics and store environment facts are taken into account.

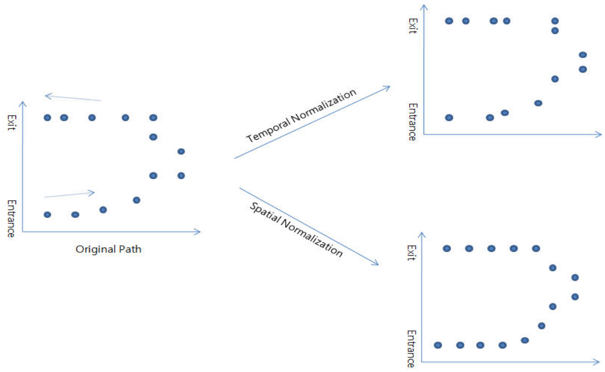

Fig. 2 Path normalization

III. Shopping Path Pattern Analysis

A. Collection of shopping data using RFID

We installed an RFID sink node (Reader) on the ceiling of the store, an RFID repeater inside the shelves and an RFID tag in the shopping cart to collect customer shopping data (Fig. 3). For identifying individual information we mapped the collected shopping data with personal information (name, age, etc).

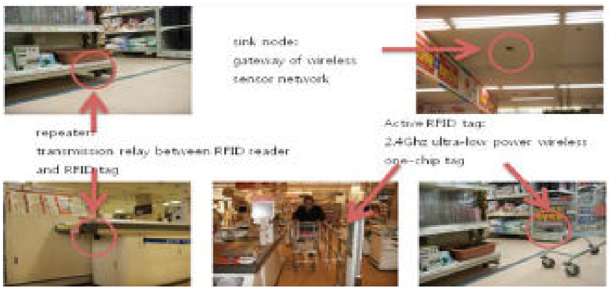

Fig. 3 RFID Device System

For more precise shopping path information, we developed an Ultra-low Power Wireless System One-Chip that uses a 2.4GHz frequency band as an active tag type. The active tags were installed in the shopping carts and a reader received data in a predefined interval. All collected data were sent to the storage server (Fig. 4).

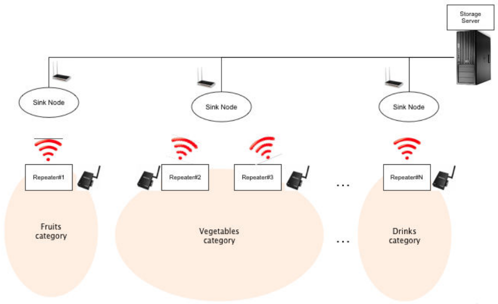

Fig. 4 RFID System Composition

B. Proposed similarity method using LCS

The basic idea for shopping path pattern identification is to extend the LCS method [6] using the newly proposed similarity between the individual customers' paths.

The LCS method finds the LCS between two sequences. If X = (x1, x2...xm), Y = (y1, y2...yn) are sequences, the LCS is,

$$LCS(X_{1\ldots i},Y_{1\ldots j})=\begin{cases} 0 & \mathit{if}\quad i=0 \,\|\, j=0 \\ LCS(X_{1\ldots i-1},Y_{1\ldots j-1}),x_i & \mathit{if}\quad x_i=y_j \\ \max(LCS(X_{1\ldots i-1},Y_{1\ldots j-1}),LCS(X_{1\ldots i-1},Y_{1\ldots j-1})) & \mathit{otherwise} \end{cases} \qquad (1)$$

Because a shopping path can be referred by node ID locations (RFID reader ID) where a customer has visited, a shopping path can be expressed with the location ID sequence in which the shopper visited. The greater the length of the common sequence between the visited paths of two customers, the greater the similarity between the two customers' shopping path pattern. If customer-1 has a shopping path of A-B-C-D-E-F, and customer-2, -3, and -4 each have shopping paths of A-B-E-Z-F, A-B-C-F, and A-B-Y-Z-F, respectively, the LCS results of customer-1 and the others are A-B-E-F, A-B-C-F, A-B-F in order (Fig. 5). The LCS of customers 1-2 and 1-3 have the same length 4 of common subsequence. In this case, LCS cannot determine which customers have a more similar shopping path pattern.

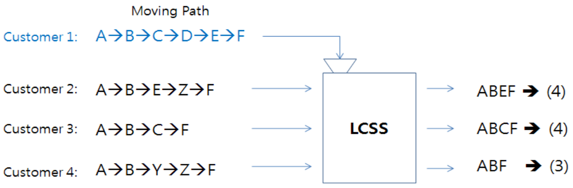

Fig. 5 Example of LCS

To solve this problem, we provide a new similarity method by extending the LCS using relative distance.

$$new\_Similarity(x,y)=\frac{LCSS(x,y)}{Length\_of\_x+Length\_of\_y} \qquad (2)$$

If we re-apply the previous examples using (2), we can find the customer who has the closest shopping path pattern. Customer-1 has a shopping path distance of 6, and customer-2 and -3 have shopping path distances of 5 and 6, respectively. By comparing 4/ (6+5) and 4/ (6+4) through (2) we can determine that customer 3 has a more similar shopping pattern to customer 1. Fig. 6 depicts this process. This means that the proposed similarity method can overcome the limitations both in the Euclidean distance measurement, which was not adequate for the store environment with many blocking obstacles, and in the sequential path.

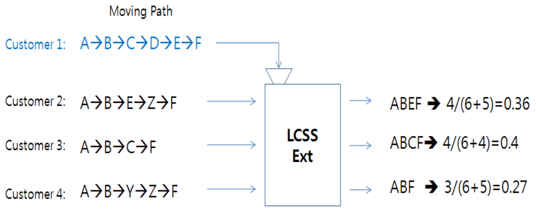

Fig. 6 Supposed similarity method

C. Main-Shopping-Path-Pattern clustering method

We developed the Main-Shopping-Path-Pattern clustering method to determine the K main shopping path in store. K is the initial clustering count such as K-means clustering. Randomly initial K paths are selected among all customer paths in order to include similar paths because each K cluster of paths is used for finding similar paths among clusters. The cluster loop is repeated until there is no change in each group or when a user-defined iterating times and then K-number of the correct LCS sequences is finally generated. (Fig. 7 and Fig. 8)

```
Main-Trajectory-Grouper (input : moving-paths, iterate-count)
    randomly selecting K-initial Trajectory in moving-paths

    loop ( stable or user determined iterate-count)
        for-each k in K
            grouping  moving-paths by (2)

        selecting first LCS in each k groups

    return k
```

Fig. 7 Pseudo-code of Main-Shopping-Path-Pattern-Grouping

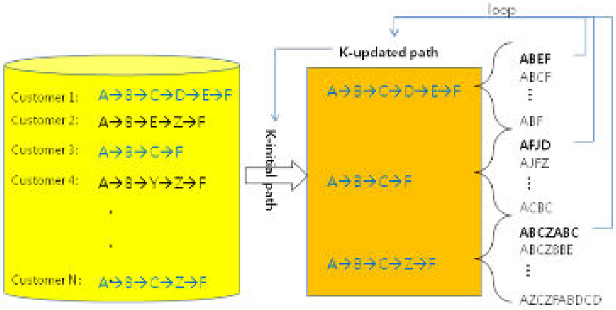

Fig. 8 Procedure of suggested method

The suggested Main-Shopping-Path-Pattern clustering method can not only detect major the customers' shopping path sequence but also identify the Hot spot and Dead spot areas. The traditional Hot spot and Dead spot identification method uses statistical analysis to sum up and compare all the locations where the shopper visited. One of the main advantages of the suggested algorithm is its ability to simultaneously identify both spots and the main shopping paths. As the LCS is characterized by grouping the main shopping path sequences in travel order, the most repeatedly appearing nodes among all sequence groups are regarded as a Hot spot, and the most rarely visited area as a Dead spot.

IV. Experimental Test And Results

To apply the proposed algorithm, we conducted a test and analyzed data for an actual large discount store located in Seoul, Korea. The store is a single floor building with an average of 554 customers daily.

More than 200 RFID Tags were installed on shopping carts and 200 readers on shelves to collect the customers' shopping traces. The data were collected for a week in February 2011 and filtered shopping paths were obtained for Monday, Wednesday and Friday.

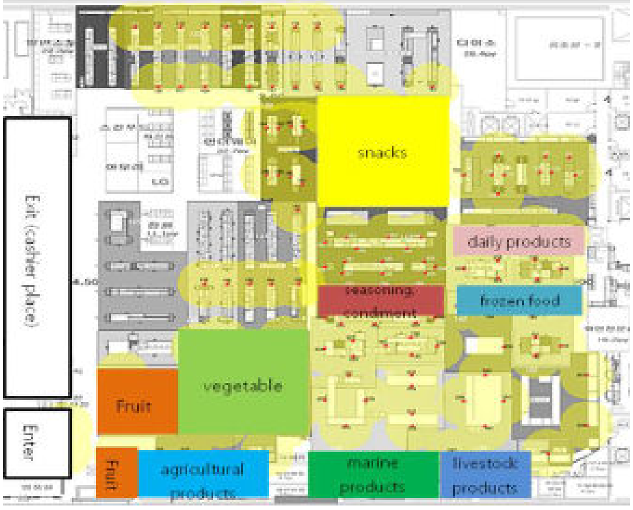

Fig. 9 Region of installed RFID devices and main products

Similar results were generated as the parameter K, the number of groups, was changed from 3 to 5. Fig. 10 depicts the result when K was set to 5 and the analysis run again. The result showed that most customers started their shopping from the entrance and mainly shopped in a counter-clockwise direction because the entrance and the cashier's counter are located in the lower left and the upper left of the picture, respectively. This layout made the customers tend to move in a counter-clockwise direction. Notably, products in the top 10 sales ranks were mostly located on the lower side of the store layout, whereas products displayed in the upper part of the layout (interior products, hobby products, car products) were rarely included in the shopping sequence and accordingly had very low actual sales.

Fig. 10 reveals that the circled areas were mostly located close to the entrance, and that the areas within 5 meters of the entrance were the first Hot Spot. Furthermore, the area before moving to the casher's counter after the shopping had been completed had the most overlapping patterns and was determined to be the second Hot Spot. Although few purchases were made in this area, it is likely to be a highly effective area for demonstration and should be used to display promotional products and hot products in order to generate more purchases. The triangle area is a bridge area that connects the first Hot Spot and the second Hot Spot, and most seasoning products and kitchen product purchases were made in this area. However, few purchases were made in the area to the left side of the triangle, even though the customers' shopping paths included this area. These results indicated that the store manager needs to change product display and promote sales through product analysis.

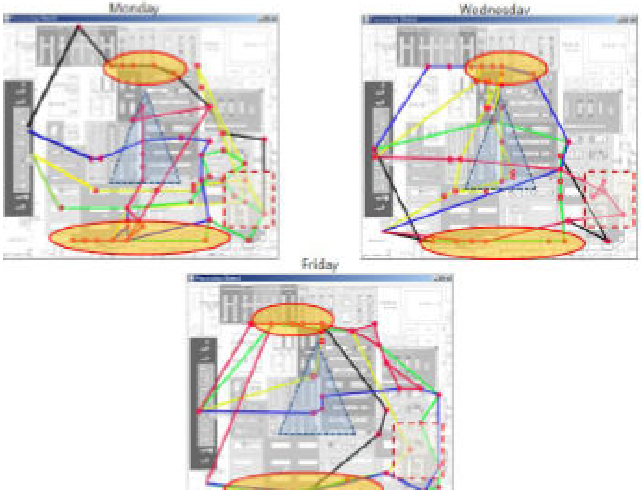

Fig. 10 Results of finding shopping movement patterns

V. Conclusion

Existing customer analysis in retail stores has relied on basket analysis or sales statistics, and has rarely included analysis for service efficiency or customer behavior pattern. However, our study provides a method to identify customers' shopping paths or major sales areas by collecting and analyzing information on customers' main travel path, which was not provided in the existing customer analysis techniques. Existing customer analysis methods using Euclidean distance suffered tumbling issues with distant spots and data processing based on the measurement. We have expanded the LCS technique and developed a new method to identify the customers' main pattern. This new method provides information necessary to decide about customers' shopping sequence and to determine meaningful spots in stores.

Based on this analysis, the results will increase understanding of the customers' consumption behavior and will assist in deciding whether product display and layout need to be changed. The proposed more quantitative method improves existing qualitative analysis which mainly relied on store employees' daily experience and provides objective numbers in order to provide high-quality services to customers and increase revenues accordingly.

For future research, we suggest combining our analysis with legacy system information such as customer purchase history in order to develop an intelligent store analysis system capable of improving operational efficiency and expanding sales. More analysis models need to be developed for more detailed analysis of the shopping behavior of diverse customers, along with the development of various measurement indexes to analyze the store environment. The present multidimensional analysis facilitated the extraction of information not previously available from existing research.

We could quantify the information on both the customer and the store by expanding the existing one-dimensional analysis into multidimensional analysis. The results reveal the need for more objective indicators in future advanced stores. We are planning to conduct more varied analysis based on those indicators. We developed new optimization technology to support decision-making on point of sale and shelf locations that will reduce customer traffic congestion, automate some of the sales processes, expand automated services to improve customer service and maximize profit for both manufacturing and business distribution. This promises to be developed into a customer service knowledge technique.

References

[1] Cox, K. (1964), The Responsiveness of Food Sales to Shelf Space Changes in Supermarkets, Journal of Marketing Research, 1(2), 63-67.

[2] Dickson, P. R. and Sawyer, A. G. (1986), Point-of-Purchase Behavior and Price Perceptions of Supermarket Shoppers, Working Paper No. 86-102, Marketing Science Institute, 1000 Massachusetts Ave., Cambridge, MA 02138.

[3] Farley, J. U. and Ring, L. W. (1996), A Stochastic Model of Supermarket Traffic Flow, OPERATIONS RESEARCH, 14(4), 555-567.

[4] Gil J., Tobari E., Lemlij M., Rose A., Penn A. (2009), The Differentiating Behaviour of Shoppers: Clustering of Individual Movement Traces in a Supermarket, Proceedings of the 7th International Space Syntax Symposium.

[5] Harris, D. H. (1958), The effect of display width in merchandising soap, Journal of Applied Psychology, 42(4), 283-284.

[6] Hirschberg, D. S. (1977), Algorithms for the longest common subsequence problem, Journal of ACM, 24(4), 664-675.

[7] Hou, J-L. and Chen, T-G. (2011), An RFID-based Shopping Service System for retailers, Advanced Engineering Informatics, 25(1), 103-115.

[8] Hoyer, W. D. (1984), An Examination of Consumer Decision Making for a Common Repeat Purchase Product, Journal of Consumer Research, 11(3), 822-829.

[9] Hui, S. K., Bradlow, E. T. and Fader, P. S. (2009), Testing Behavioral Hypotheses Using an Integrated Model of Grocery Store Shopping path and purchase Behavior, Journal of consumer research, 36, 478-493.

[10] Hui, S. K., Fader, P. S. and Bradlow, E. T. (2009), Path Data in Marketing: An Integrative Framework and Prospectus for Model Building, Marketing Science, 28(2), 320-335.

[11] Larson J. S., Bradlow E. T. and Fader P. S. (2005), An exploratory look at supermarket shopping paths, International Journal of Research in Marketing, 22(4), 395– 414.

[12] McClure, P. J. and West, E. J. (1969), Sales Effects of a New Counter Display, Journal of Advertising Research, 9, 29-34.

[13] Newman, A. J., Yu, D. K. C. and Oulton , D. P. (2002), New insights into retail space and format planning from customer-tracking data, Journal of Retailing and Consumer Services, 9(5), 253-258.

[14] Uotila, V. and Skogster, P. (2007), Space management in a DIY store analyzing consumer shopping paths with data-tracking devices, Facilities, 25(9), 363-374.

In-Chul Jung graduated from City liberal arts(BS), Incheon University in 2002 and MS from Industrial Systems Engineering, Dongguk University, 2005 and Ph.D student in Industrial Systems Engineering, Dongguk University from 2006. Interests include machine learning, data mining, agents, and intelligent information systems

Young S. Kwon graduated from Seoul National University, Industrial Engineering (BS), 1978 and MS from Industrial Engineering, Korea Advanced Institute of Science and Technology, 1981 and Ph.D. from Korea Advanced Institute of Science and Technology graduate,1996. He has worked professor in Dongguk University, Industrial and Systems Engineering since 1981. Interests include data mining, intelligent information systems
.