A Sequence Pattern Matching Approach to Shopping Path Clustering

In-Chul Jung and Young S. Kwon
Department of Industrial and Systems Engineering
Dongguk University
South Korea

Yung-Seop Lee
Department of Statistics
Dongguk University
South Korea

Abstract

We can have a new perspective about customer by analyzing customer shopping path data because customer should go to a specific place to buy some product in store. The analysis of path data is time-consuming work. In this way, the commonly used method is clustering algorithm in order to understand the tendency of data. But a general clustering algorithm is not suitable to identify the shopping path of customers because of various spatial constraints such as aisle layout or other physical obstructions in the store. In this paper, we propose a shopping path clustering algorithm with sequence pattern matching method, LCSS (longest common subsequence) method, which groups similar moving path for shop in order to understand characteristics of shopping. Experimental results using real data obtained from a grocery in Seoul 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

1. 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 discount stores has been only low-price strategy with the expansion of many branch stores. Recently, however, they have struggled with the decreased consumer spending due to the economic recession. They should have changed the only low-price strategy to new marketing strategies such as aggressive promotions to customers by this situation. They run a market basket analysis or regional analysis based on customer purchase history and demographic information for new strategy. For example, we can recommend products which given customers are interested by attempting to analyze a purchase history and customers segmentation. Or we can provide a new specific customer service by attempting to analyze and identify demographic information and regional analysis. If there is more information available about customer, we can understand our customer. We can appeal to the customer by understanding our customer preference and behavior like shopping path. Recording and understanding the behavior of customers is paramount and a key factor influencing the success of any retail business.

According to the research of Newman et al. (2002) on a new methodological approach to analyzing in-store customer behavior with a view to optimizing space and store performance, understanding these customer processes and movement patterns thus helped the retail collaborator maximize the performance of the store. If we can find the areas where most sales activities occur and where customers tend to stay for a long time in the store, store manager can understand 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.

In this paper, therefore, we propose a new method of shopping path clustering algorithm which groups similar moving path for shop in order to understand customer shopping behavior. We developed a new similarity

measure of moving trajectory by extending the longest common subsequence (LCSS) method which is one of sequence pattern matching methods. Our proposed method can find the main shopping paths that are capable of identifying the hotspots where most of the customers' visits are made and the dead spots with few visits. So we can understand shopper behavior in a store. We applied the proposed method to analyze real dataset of a grocery supermarket store in Seoul, as case study, to demonstrate the advantages and usefulness of the new method.

2. Related Literature

Some researchers have studied customer behaviors using direct observation or questionnaires. In-store advertisements and promotions have proven records to amplify the magnitude of unplanned purchasing among consumers (McClure and West, 1969). Cox (1964) measured relationship between shelf space and product sales. Dickson and Sawyer (1986) investigated consumers' knowledge and use of price information at the supermarket point of purchase (POP) and Hoyer(1984) provided a view of decision making based on the notion that consumers are not motivated to engage in a great deal of in-store decision making at the time of purchase when the product is purchased repeatedly and is relatively unimportant.

Farley and Ring (1996) recorded the movement of some customers by following them in order to analyze the shopping path as one of the shopping behaviors. 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. In recent years, technological advances such as inexpensive RFID and GPS and etc. have been applied to the analysis of customer behavior (Larson et al., 2005; Uotila and Skogster, 2007; Gil et al., 2009; Hou and Chen, 2011).

Larson et al.(2005), Hui et al.(2009) and Gil et al. (2009) have tried to solve these problems by using radio-frequency identification (RFID) and clustering techniques to analyze customers shopping path. Many research mainly have used to clustering algorithm among data mining methods to detect the main shopping path patterns. Larson et al. (2005) and Hui et al. (2009) tried to identify the major shopping path using the k medoids clustering algorithm. However, a clustering algorithms using euclidean distance similarity measure on 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 walk 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.

For example, as shown in Figure 1, if we measure the euclidean similarity 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 clustering algorithms using euclidean similarity measure are inadequate for in-store shopping path pattern grouping.

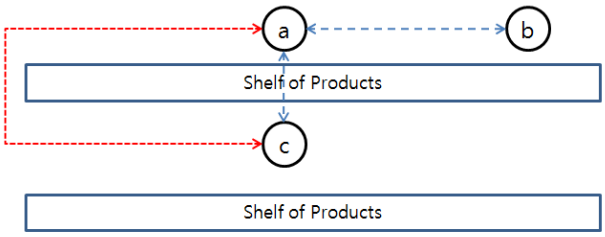

Figure 1: Euclidean distance and actual customer path in a store

Furthermore, the attribute's length of all objects for clustering algorithm have to be same length. But customer shopping path lengths are 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 paths 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. For example, when a clustering algorithm like K-mediods is used, the input attributes have to be identical because of the characteristic of the algorithm.

Therefore, different customer shopping length needs to be converted into an same length, and the normalization is applied to equalize the size of the trace. As in figure 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 paper provides a method for detecting main customer shopping path patterns in which all these path characteristics and store environment facts are taken into account.

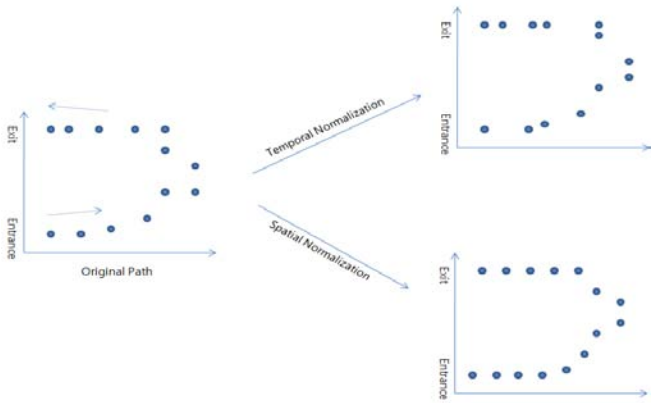

Figure 2: Path normalization

3. Shopping Path Pattern Analysis

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 clustering algorithm that is capable of identifying the hotspots where most of the customers' visits are made and the dead spots with few visits.

3.1 Proposed similarity measure using LCS

The basic idea for shopping path trajectory similarity is to extend the LCS method (Hirschberg, D. S., 1977). The LCS problem is to find the longest subsequence common to all sequences in a set of sequences. If X = (x1, x2...xm), Y = (y1, y2...yn) are sequences, the LCS is :

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

Because movement trajectory can be referred by sequences of locations where a customer has visited, we can define a shopping path with sequence of the location IDs which a shopper visited. The longer the length of LCSS between shopping paths of two customers is longer, more similar trajectory of two customers. For example, customer 1 has a shopping path of <A-B-C-D-E-F>. Each customer 2, customer 3 and customer 4 have shopping path 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 <ABEF>, <ABCF>, <ABF> in order (Figure 3).

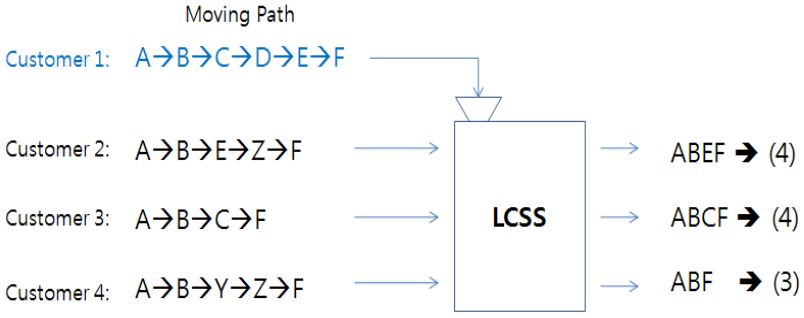

Figure 3: Example of LCS

But, the LCS(customers 1, customers 2) and LCS(customers 1, customers 3) have the same length 4 of common subsequences. In this case, origin LCS cannot determine which customers have more similar shopping path pattern. To solve this problem, we suggest a new similarity measure by extending the origin LCS like :

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

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 path length of 6 and customer 2, customer 3 have path lengths of 5 and 6, respectively. By comparing new-similarity(customer1, customer2) and new-similarity (customer1, customer 3) using (2) we can determine that customer 3 has a more similar shopping path pattern to customer 1. Figure 4 depicts this process.

The proposed similarity measure based on LCS which is a better measure for moving paths that have different moving length compared to Euclid distance and so on.

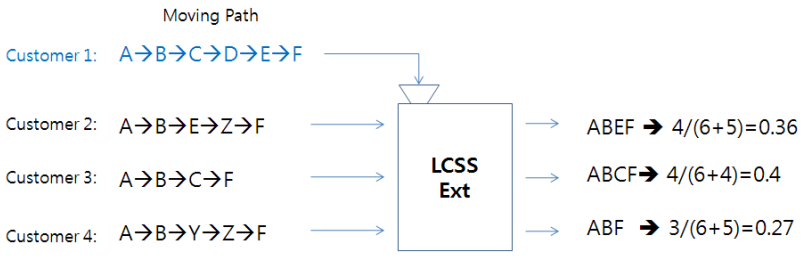

Figure 4: Process of supposed similarity measure

3.2 Shopping Path Clustering Algorithm

We developed the shopping path clustering algorithm using (2) to determine k shopping path in store. k is the initial clustering count such as k-means clustering. Figure 5 is pseudo code of proposed shopping path clustering. First, initial k moving paths are randomly selected among all paths in order to find k clusters. Second, find the most similar moving path with each clusters according to arg_max_k new-sim (moving path, cluster K). The current moving path is inserted to k-th cluster.

```
이름:  K-Moving Path Clustering
입력:  moving-paths, K, iterate-count
출력:  cluster_result

1     randomly selecting initial K-groups in moving-paths
2     for-each moving-path in moving-paths {
3        for-each k in K-groups {
4           cluster_index = arg max_k (LCSS-SIM(moving-path, k))
5                 insert moving-path to cluster_result (cluster_index)
6        }
7     }
8     return cluster_result
```

Figure 5: Pseudo-code of Shopping Path Grouping

Figure 6 describes 5 moving paths on 5x5 cell map for explanation about procedure of proposed algorithm. All moving path of objects are O1 = <1,2, 3, 4, 5, 10, 15, 20, 25, 24, 23, 22, 21>, O2= <1, 2, 3, 8, 7, 6>, O3= <1, 2, 3, 4, 9, 14, 13, 12, 17, 18, 19, 24, 23, 22, 21>, O4= <1, 2, 7, 11, 16, 21> , O5=<1, 6, 11, 16, 21>. All objects start from 1 cell to 6 or 21 cell. For simplicity, we assume that k is 2.

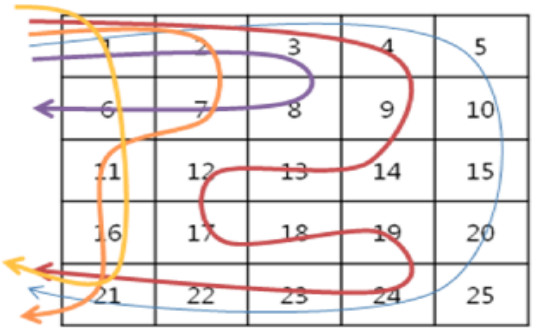

Figure 6: Moving paths

First, our clustering algorithm randomly chooses k moving paths from the total moving paths and uses these as the initial clusters. If moving paths of O2 and O2 are selected:

O1= <1, 2, 3, 4, 5, 10, 15, 20, 25, 24, 23, 22, 21>
O2= <1, 2, 3, 8, 7, 6>
O3= <1, 2, 3, 4, 9, 14, 13, 12, 17, 18, 19, 24, 23, 22, 21>
O4= <1, 2, 7, 11, 16, 21>
O5=<1, 6, 11, 16, 21>

Second, our method finds moving path which are the most smilar to cluster of O2 using (2). O4 is most similar. O2 and O4 is same cluster.

O1= 3/(13+6) = 0.16
O3= 3/(15+6) = 0.14
O4= 2/(6+6) = 0.17

Next, our method finds moving path which are the most smilar to cluster of O5 using (2). O1 is most similar. O1 is most similar. O5 and O1 is same cluster.

O1= 2/(13+5) = 0.11
O3= 2/(15+5) = 0.1

Finally, O1 is same cluster with O2 and O4 because LCSS-SIM is 0.14 > 0.1 :

LCSS-SIM(O2, O3) = 0.14
LCSS-SIM(O5, O3) = 0.1

Figure 7 show the final clustering result. We can analyse a result of clustering that there are two moving patterns. First pattern has a characteristic to show long moving length and mainly visiting cells are {1, 2, 3, 4, 5, 21, 22, 23, 24, 25}. The other pattern has a characteristic to show short moving length than first cluster and mainly visit cells are {1, 2, 3}.

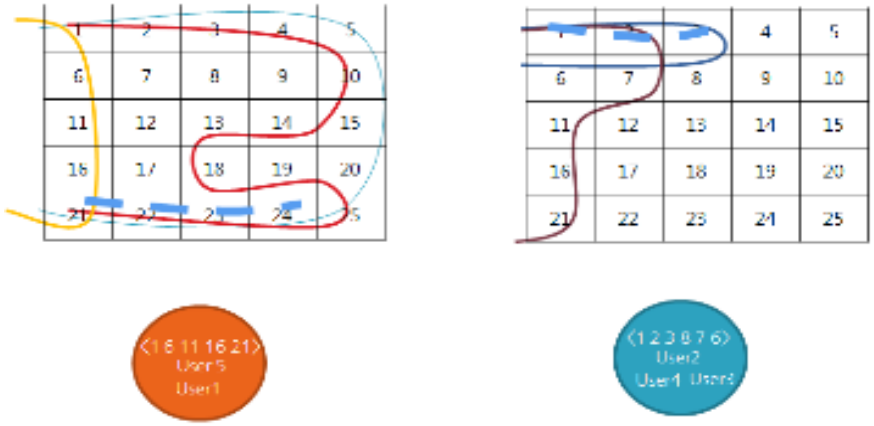

Figure 7: Moving paths

The proposed shopping path clustering algorithm can not only detect major the customers' shopping paths but also simultaneously identify the hot spot and dead spot areas. As the LCS is characterized by grouping the main shopping paths 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.

4. Case study

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.

4.1 Collection of shopping data using RFID

We installed an RFID sink node (RFID Reader) on the ceiling of the store, an RFID repeater inside the shelves like Figure 8. For identifying individual information we mapped the collected shopping paths with personal information (name, age, etc).

Figure 8: 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 transmitted a signal to a RFID repeater in a predefined interval. And then the RFID repeater transmitted location data to RFID readers. All collected data were sent to the storage server like Figure 9.

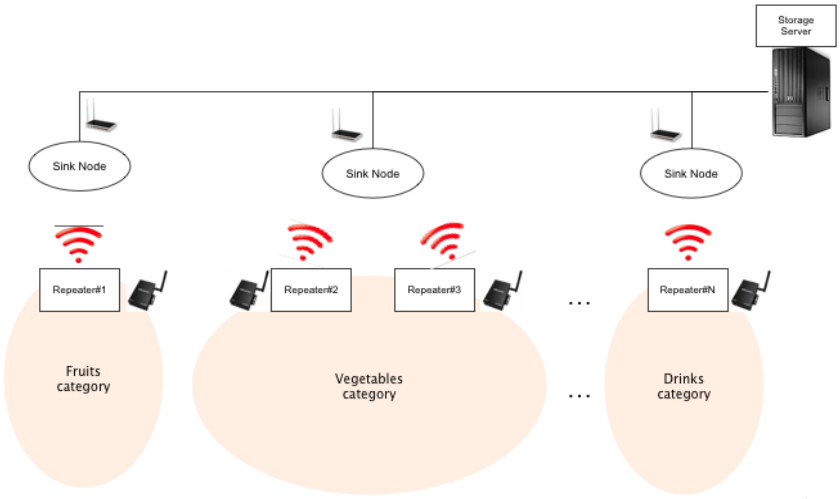

Figure 9: RFID System Composition

4.2 Result

For analyzing cluster results, we changed the parameter k, cluster count, from 3 to 5. But similar results were generated that figure 10 depicts the clustering 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.

.

Figure 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.

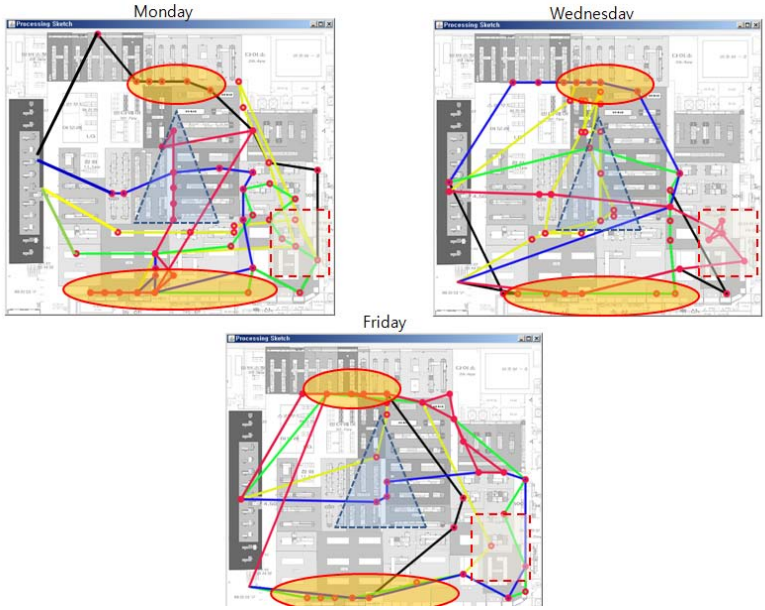

Figure 10: Results of finding shopping movement patterns

5. 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.

By understanding the components of the path, you can better strategize ways to maximize a product's impact upon the consumer at each phase, and ultimately win the sale.

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.