International Journal of Advanced Science and Technology
Vol.116 (2018), pp.49-58
http://dx.doi.org/10.14257/ijast.2018.116.05

A Study on Improvement of Algorithm for Measuring Similarity of Consumer Movement Paths in Shopping Mall

Myeong Bae Kim[1], Sung Won Jung[2], Yong Duk Chi[3] and Gwang Yong Gim[4*]

[123]*Department of IT Policy Management, Soongsil University*
[4]*Department of Business Administration, Soongsil University*
[1]*Mbkim77@gmail.com*, [2]*jdate@naver.com*, [3]*ofcyd@daum.net*, [4]*gygim@ssu.ac.kr*

Abstract

In this paper, we proposed an improved CMPSI (Consumer Moving Path Similarity Index) algorithm to apply the consumer characteristics of shopping malls and beacon data. The proposed algorithm is a standardized index for defining the standardized similarity considering shopping route and stay time in each beacon area and combining them with similarity considering the definition. The improved CMPSI proposed in this paper is expected to be used as important information for marketing activities to increase profit by combining with consumer purchase information by using beacon service and to apply to other major offline shopping mall area.

Keywords: *Omni-channel, B2C commercial area, Beacon, Moving Path Similarity Index*

1. Introduction

Wikipedia says "Omni-channel is a cross-channel business model that companies use to improve their customer experience. The approach has applications in healthcare, government, financial services, retail and telecommunications industries, and includes channels such as physical locations, FAQ webpages, social media, live web chats, mobile applications and telephone communication. Companies that use omni-channel contend that a customer values the ability to be in constant contact with a company through multiple avenues at the same time." [11].

In the multi-channel environment, if each channel operates independently and competition is on-line and off-line, omni-channel has a characteristic of on-line and off-line being a win-win relation together with customer-oriented organic channel operation [10].

This paradigm shift means that companies can expand the scope and method of existing marketing activities. Particularly in the case of an offline shopping mall, a more efficient marketing activity becomes possible because the customer's action area can know the information on the online/mobile in a situation where the market position gets narrowed.

In order for omni-channel service to be available, it means that the integration of the online/mobile customer and offline customer identification ID, and the code for the product should be integrated with each other.

From the perspective of a marketer analyzing a customer, they will want to know the characteristics of their customers. Beacon technology is one of the technologies that can collect such information.

Thus this study is going to verify the practicality of this method by suggesting the similarity algorithm considering the travel route and the time required for practical marketing activities in large-scale offline stores and applying it to actual data. In addition, this study is going to suggest a method for analyzing the data collected by BLE Beacon for

Received (January 4, 2018), Review Result (March 8, 2018), Accepted (March 12, 2018)
* Corresponding Author

ISSN: 2005-4238 IJAST

effective marketing activities in the offline store, and propose the fields that can be extended.

2. Related Work

2.1. The Change of Shopping Characteristics

In paper [7], the author Yeom Minseon proved that showrooming consumers have reasonably and deliberately shopping behavior to obtain economic benefits by complementing differentiated advantages of each shopping channel.

In paper [9], the author Choi Hyunseung & Yang Seongbyeong mentioned that Omni-Chanil consumers are changing from past showrooming shopping to Web rooming. They say this is due to increased use of smart devices and changes to omni-channel.

In paper [5], the author examined the various users of omni-channel service in Korea. They also deducted “the initiative for selecting shopping time and places, eliminating redundant and unnecessary work, doubling discount and accumulation benefits through affiliate/channel integration, and a close on/offline linkage through systematized processes” as for experience factor of omni-channel service that influences on improvement of customers‘ shopping. As a means to improve these empirical factors, customer's location information can be used as very crucial data.

2.2. Beacon Practicality

"Beacon is a location- based service of wireless communication technology that transmits and receives various information such as GPS information in connection with BLE signal based on Bluetooth low energy (BLE) communication protocol," mentioned by Kim Suyeon *et al.*, in [5].

In paper [8], the author Oh Amseok suggested "smart factory logistics management system equipped with intelligence provision the function, such as providing of emergency notification, road sign information, parking information and smart logistics tracking function" using beacon. The principle of this system is to track the moving route and storage location based on the signal strength of the beacon tag.

In paper [4], the author Kwon Daewon suggested a system for monitoring the indoor position of demented patients using beacons in order to prevent the disappearance of demented patients who may be roaming. He applied a reversible method in which a dementia patient wears a beacon wearable device and detects a signal in a beacon signal receiving device installed in a specific place.

2.3. Similarity Algorithm

In the existing research, various types of clustering methods have been studied, which results in the problem of how to measure the similarity between individuals. Among them, this study is going to discuss related studies on clustering algorithms with characteristics that occur over time.

In paper [14], the author Larson et al. used the k-medoids algorithm to cluster with customer moving path data in an offline shopping mall. This is a clustering method using Euclid-based distance functions.

In paper [6], the author Yang, Seungjoon *et al.*, suggested a distance-index matrix calculation method between moving points, and used a method to calculate the shortest path in advance by expressing movable positions as nodes, and directly connecting nodes that have no obstacle to movement.

As a clustering method based on the center point, these methods have a disadvantage that the movement data is deformed, because the moving path data are normalized to the same length for the same time in [2].

As for a method to cluster the density basis there is Density Based Spatial Clustering of Applications with Noise in [12]. This is a method of clustering with the epsilon centered around a specific object and the number of individuals within that radius.

In paper [15], the author Tung *et al.*, suggested COD-CLARANS (Clustering with Obstructed Distance-CLARANS) which considers obstacles in space. It is based on CLARANS, a partitioned clustering algorithm, which extracts a sample from a large volume of original data and finds the cluster center. Here, the length of the shortest path that bypasses the obstacle is used as a distance (dissimilarity).

The above methods are not suitable for the method of measuring the similarity of the entire moving path because of the feature of clustering based on the positions of the objects on the coordinates. Sequences should be considered to compensate for this.

In paper [12], the author Hirschberg, D. S. suggested LCSS (Longest Common Subsequence) function to measure the similarity between character strings. This method finds the longest sequence among the partial sequences common to both strings of sequences.

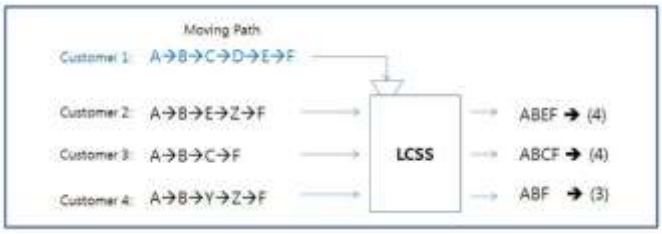

Figure 1. Measurement Example of LCSS

This method has an advantage that it can be used even if the length of the character string is different, but there is a characteristic that the longer the length of the comparison string, the higher the value. When we apply this method to the customer's moving path in the offline space, a couple of problems arise. It is that it is highly likely that the LCSS length will be large among the individuals having a long travel length. In paper [2], the author Jung In-cheol suggested LCSS_SIM function that standardized LCSS values to overcome these shortcomings.

It is to correct the total length of travel that the two objects moved to the length of LCSS length. The similarity function of that changed LCSS is as follows.

$$relative\_lcss(x,y) = \frac{lcss(x,y)}{length(x)+lengh(y)} \quad (1)$$

Here, *length(x)* is the travel distance of x. *relative_lcss* has an advantage if it is possible to make a relative comparison considering the entire travel distance.

Here, the range of the value of *relative_lcss* is 0~0.5. In order to have min-max standardization of this to a value between 0 and 1, this study suggested *relative_LCSS_SIM* as follows.

$$relative\_lcss(x,y) = \frac{lcss(x,y)}{length(x)+lengh(y)-lcss(x,y)} \quad (2)$$

However, *relative_LCSS_SIM* does not exactly have a value between 0 and 1. If x and y have the same length and all match, then the denominator becomes 0 and calculation becomes impossible. In addition, there is a disadvantage in that no real time travel time is taken into consideration due to obstacles in the offline store or a residence time at the location.

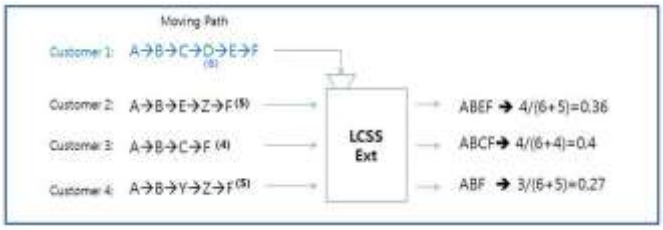

Figure 2. Measurement Example of relative LCSS

In paper [3], the author Kang Hyeyoung *et al.*, suggested a similarity algorithm considering the Common Visit Time Interval (CVTI) under the cell-space basis. This study suggested common visit time between object a and b as follows.

$$CVTI(a,b) = \sum_{i=1,j=1}^{n,m} |a_i \cdot I \cap b_j \cdot I| \quad (3)$$
$$where\ a_i \cdot c = b_i \cdot c$$

Here, $a_i \cdot I$ refers to the visit time interval of the ith travel position of a object, and $a_i \cdot I \cap b_j \cdot I$ the common time in the ith position of a object and jth position of the b object.

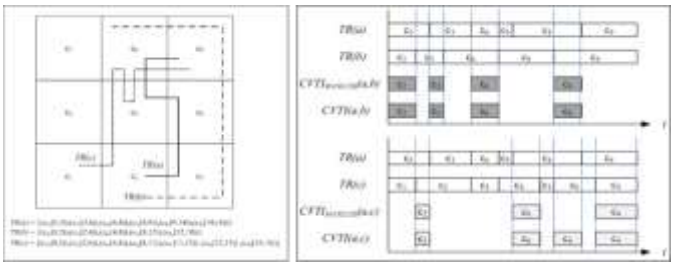

Figure 3. Measurement Example of CVTI

This study also proposed an index to measure similarity between trajectories by adding a time concept to LCSS (Longest Common Subsequence) method. The $CVTI_{LCSS}$ is defined as follows.

$$CVTI_{LCSS}(S_{LCSS}(a,b)) = \sum_{(c,i,j)\epsilon S_{LCSS}(a,b)} (a_i \cdot I \cap b_j \cdot I) \quad (4)$$

Here, $CVTI_{LCSS}$ considering only the cells appearing in CVTI, LCSS and S_{LCSS} denotes the length of the longest common substring, and a set of all longest common substring lengths is S.

Also, $CVTI_{LCSS}$ finally took the maximum value and defined $CVTI_{MAXLCSS}$ as follows.

$$CVTI_{MAXLCSS}(a,b) = \max(S_{LCSS}\ \epsilon\ S \mid CVTI_{LCSS}(S_{LCSS})) \quad (5)$$

$CVTI_{MAXLCSS}(a,b)$ has the advantage of considering the visit time of the offline shopping mall, but there is a limitation that only the section having the same elapsed time since the first entry is considered.

Preliminary studies have had problems in applying to offline shopping malls. Also the matter that should be considered for similarity measure is as follows. First, shopping time is different for all customers. Second, the time spent visiting the zone should be considered even if the elapsed time of the visit does not match. Third, the similarity should be standardized to a value between 0 and 1, since it must be comparable to the objects in other time.

3. Proposed Model

3.1. Improvement of Algorithm Considering Travel Route

The existing *relative_LCSS_SIM* does not have a value between 0 and 1 exactly. Thus, this study use *MIN(length(x), length(y))* in denominator for correct standardization. That is, the minimum value of the number of visited paths of two objects is used.

$$S_{LCSS(x,y)=}\begin{cases}0 & if\ i=j=0\\ c_{i-1,j-1} & if\ i,j>0\ and\ x_i=y_j\\ MAX\{c_{i-1,j}\ , c_{i,j-1}\}\ if\ i,j>0\ and\ x_i\neq y_j\end{cases} \quad (6)$$

Here, in order to *standardized-relative_LCSS* with the value between 0 and 1, when using the minimum value in the denominator, it can be expressed as follows.

$$Std-relative\_S_{LCSS}=\frac{S_{LCSS(x,y)}}{Min(length(x),lengh(y))} \quad (7)$$

Here, $S_{LCSS(x,y)}$is the longest common substring of object x and y, and $Min(length(x), lengh(y))$ is the minimum value of travel route of x and y. Proposed $Std-relative\_S_{LCSS}$ has a value between o and 1 as $S_{LCSS(x,y)}$ is always less than or equivalent to the value of $Min(length(x),\ lengh(y))$.

3.2. Improvement of Algorithm Considering Stay Time

The stay time means the time that stayed in the zone where the beacon is installed. The beacon zone is an area configured with a predetermined radius, and records the time at which the customer's mobile enters and the time the customer departs the area. Departure Time - Entry time is defined as Stay Time.

The stay time similarity is defined in consideration of the stay time of the zone included in *LCSS(x,y)*, regardless of the elapsed time from the initial entry time of the shopping mall based on the characteristics of shopping. This can be expressed as follows.

$$adj-CVTI(x,y)=\{\sum_{i=1,j=1}^{n,m}|x_i\cdot I\cap y_j\cdot I|\}/MIN(T_x,T_y) \quad (8)$$

Here, $\sum_{i=1,j=1}^{n,m}|x_i\cdot I\cap y_j\cdot I|$ is the common stay time of the same zone of objects x and y, and $MIN(T_x, T_y)$ is the minimum value of the total shopping time of a and b objects. Since the total sum of the time the two objects stayed in common is less than or equal to the minimum value of the total shopping time of the two objects $adj-CVTI(x,y)$ is exactly between 0 and 1.

But, some restrictions are required. There can be two possible ways to re-enter/depart the same zone repeatedly. It is the method of applying the entire stay time of the zone and considering the average stay time. This will be available as an option. In addition, the applicable time can be replaced by not only the stay time but also the travel time between zones, but data may be distorted if there is a toilet or rest space between zones.

3.3. Mixed Similarity Algorithm

This study proposed a customer moving path similarity measurement algorithm for a large off-line shopping mall by combining *standardized-relative_LCSS* and *adj_CVTI* suggested in previous chapter. Both values have a value between 0 and 1. Therefore, this study propose a *Consumer Moving Path Similarity Index (CMPSI)* algorithm with linear combination of two successive similarities. Here, if the sum of the weights of two arithmetic

formula is set to 1, the CMPSI is also an index having a value between 0 and 1. This can be expressed as follows.

$$CMPSI = \alpha(Std - relative_{S_{LCSS}}) + \beta(adj - CVTI(x,y))$$

$$= \alpha \frac{S_{LCSS(x,y)}}{Min(length(x),lengh(y))} + \beta\{\sum_{i=1,j=1}^{n,m}|x_i \cdot I \cap y_j \cdot I|\}/MIN(T_x, T_y) \quad (9)$$

Where, $S_{LCSS(x,y)}$ is the length of the longest common substring of object x and y. $Min(length(x),\ lengh(y))$ is the minimum value of the number of the travel routes of x and y.

$\sum_{i=1,j=1}^{n,m}|x_i \cdot I \cap y_j \cdot I|$ is common stay time of the same zone of object x and y. $MIN(T_x, T_y)$ is the minimum value of total shopping time of object x and y. α, β is as for weighted value of each $\alpha+\beta=1$, $0 \leq \alpha, \beta \leq 1$.

Through this algorithm, this study can measure similarity based on customer's moving path, which cannot be measured in marketing analysis area using existing beacon data, and it is expected to be able to expand the analysis area for customers.

4. Empirical Analysis

This empirical analysis used data collected from beacons installed in a large scale of offline shopping mall to analyze the spatial information and to calculate the time of the customer, the time of stay, the hub zone, and the CMPSI.

4.1. Customer Moving Path Similarity (CMPSI) Analysis Results

The customer moving path similarity (CMPSI) proposed in this study was calculated by the following procedure. Basic search was performed with the calculated information.

Table 1. CMPSI Calculation Procedures

Subject: Customer moving path Similarity (CMPSI) Input: USER, zone, entry, visual log data Output : std-LCSS, adj-CVTI, CMPSI
1. Loading Data - read.table 2. USER, hourly data alignment - order (USER, SEQ) 3. Create USER list - unique (USER) 4. ZONE sequence string creation by USER - aggregate () 5. Definition of Pattern search function considering sequence 6. USER Matrix creation 7. Create output data.frame to get the combination and pattern - data.frame () 8. Standardized LCSS value calculation length (LCSS) / (max (length (tj1), length (tj2))) 9. Calibrated visit time calculation 10. Customer moving path similarity Output 11. Export final data - write

4.2. Comparison of Similarity Index

Table 2 shows the characteristics of males who did not show characteristics by age group. For males, the moving path similarity of groups with the same sex and age group is high for all ages. On the other hand, in the case of women, the same sex and age groups have high similarity only in some the 40s and 20s, and the rest have high similarity among the sex/age groups. This could be the result deducted because the patterns of shopping are simple for men but they are very diverse for women.

These results show that the newly proposed algorithm is worthy of exploiting the moving path characteristics of customers.

Table 2. Similarity Comparision by Gender, Age

Gender	Age	Gender/Age Same	adj_Jcss average	adj_CVTI average	CMPSI average	Number of comparison customers
Male	20s	Same	0.341	0.239	0.290	10
		Different	0.281	0.222	0.251	233
	30s	Same	0.277	0.219	0.248	152
		Different	0.261	0.228	0.245	853
	40s	Same	0.352	0.276	0.314	9
		Different	0.334	0.265	0.299	352
	50s	Same	0.500	0.317	0.409	1
		Different	0.385	0.308	0.346	76
Female	20s	Same	0.287	0.226	0.257	100
		Different	0.276	0.238	0.257	721
	30s	Same	0.248	0.229	0.238	529
		Different	0.289	0.261	0.275	1,473
	40s	Same	0.334	0.280	0.307	274
		Different	0.286	0.245	0.265	1,135
	50s	Same	0.266	0.194	0.230	44
		Different	0.280	0.246	0.263	279

4.3. Moving Path based Customers Targeting

Let's take advantage of this similarity in cases when a new product is put on a specific brand store, or when planning an event space such as a display in a specific zone. Select the moving path for the purpose of the campaign and include the information in the analysis data. Figure 4 shows the data entered through the 'User Input' node in SPSS Modeler. This study selected only USER = 'standard' in the final result after inserting this object into the analysis data before applying the algorithm and calculating the degree of similarity.

Figure 4. Reference Data Input Screen in IBM SPSS Modeler

Table 3 shows the Top 10, which has high similarity based on 'standard'. When increasing the number of choices according to the purpose, it is possible to extract the target customers, and extract if you give the condition with same gender and age.

Table 3. Standard moving path and top 10 similarities

USER_1	USER_2	CMPSI	Gender	Age
standard	384873	0.868	Male	40s
standard	424329	0.868	Female	30s
standard	324744	0.865	Male	40s
standard	550201	0.834	Female	40s
standard	684128	0.834	Male	30s
standard	592334	0.834	Male	30s
standard	588895	0.818	Female	50s
standard	550201	0.818	Male	40s
standard	365839	0.814	Female	40s
standard	587860	0.814	Female	30s

4.4. Customer Marketing Using the Herb Zone

First, the herb zone selected in the analysis is the first floor entrance 1, the first floor entrance 2, the food 2, the food 3, the second floor entrance, the kids 2 zone, and all the remaining zones except the kids 2 zone are located on the first floor of the store. This means that the first floor of the store has more customers moving than the second floor, and that it is the point connecting the second floor and the first floor of the store.

If you engage in marketing activities at these locations, it will be highly likely that most customers are exposed to marketing activities.

4.5. Marketing using Customer Moving Path Similarity

As for the method to use customer moving path similarity (CMPSI), the first method is to select moving path in advance that matches the marketing strategy and target the top 100 people that are most similar to such moving path. This similarity information will be useful in cases that a new product is put on a specific brand store, or when planning an event space such as a display in a specific zone.

The second method is to extract customers who have recently purchased a specific product (such as a customer who bought a new product bag) and target the top 10 customers who have similar moving paths. Of course, in addition to the customer's moving path viewpoint, it is necessary to consider the past purchase history or life style.

Third, shopping moving path can be used to extract very unique customers from other customers having different moving path. For example, a customer with a similarity of less than 0.1 with other customers means that the customer is moving in a different pattern of moving path than other visiting customers.

It is meaningful that these three methods can be utilized as additional information from the viewpoint of moving path in the offline environment when analyzing the characteristics of customers.

5. Conclusion

In this paper, we suggested an improved algorithm after pointing out the inappropriate part of the similarity algorithm suggested in existing studies, from the viewpoint of the customer moving path of the large offline shopping mall. Also, this study applied to the beacon data using the data occurring from beacon, and as a result, it could be utilized as a customer identification index that can be used for marketing. If we identify the moving path of the visiting customers in the large commercial space and find the characteristic of each zone based on this, and apply appropriate marketing to the characteristic, it will be possible for customers to expect a more convenient space and for companies to expect profit increase.

6. Future Research Tasks

This study has some limitations. First, this study did not consider the various times when calculating the customer moving path similarity (CMPSI) value. For example, it is not appropriate to use the travel time between zones and to use the average or maximum value of the stay time of the same zone.

Second, the location of beacon installed in a large offline shopping mall applied for empirical analysis was rather not suitable in terms of utilization. A beacon was installed on the outskirts of the zones consisting of a circular main passage and a small passage in the center. This was somewhat lacking in the purpose of customized customer analysis.

Third, the customer moving path similarity (CMPSI) suggested in this study was not applied to actual marketing and neither its effectiveness was verified. There is a need for research that provides actual offers, measures customer responses, and compares them with existing marketing responses.

Fourth, it is necessary to verify whether generalization is possible for other location-based data. A variety of refining method should be devised because there will be specific matters depending on how to operate location-based data such as beacon.

Fifth, further study is needed to verify practicality by applying to the fields requiring the differentiated service according to customer's moving path.

Lastly, it is necessary to study more precisely the products of interest of customers by analyzing the linkage of stores that have a long stay time in offline by combining online purchase details and products of interest.

References

[1] M. B. Kim, "A Study on the Improvement of Algorithm in Consumer Moving Path Similarity Measurement Using Beacon Data", Soongsil University Doctoral Thesis, **(2017)**.
[2] I. C. Jung, "A Study on moving path Data Clustering Techniques for Moving Objects: Case Study: Moving path Analysis of Large Distribution Store Customers", Dongguk University Doctoral Thesis, **(2012)**.
[3] H. Y. Kang, J. S. Kim, J. R. Hwang and K. J. Lee, "Similarity Measure of Moving Object Trajectories in Cellular Space", Journal of Korea Geographic Information Systems, vol. 16, no. 3, **(2008)**, pp. 291-301.
[4] D. W. Kwon, "A study on indoor location based dementia patient monitoring system using Bluetooth beacon", Digital Convergence Research, vol. 14, no. 2, **(2016)**, pp. 217-225.
[5] S. Y. Kim, G. Y. Park, H. M. Koo and S. W. Kim, "An Exploratory Case Study on the Improvement of Omni Channel Service Type and Shopping Experience - Focusing on the Three Domestic Major Distribution Centers (Lotte, Shinsegae, Homeplus)", Design Convergence Research, vol. 15, no. 5, **(2016)**, pp. 85-103.
[6] S. J. Yang, I. C. Jung, and Y. S. Kwon, "Customer shopping moving path pattern analysis using RFID data", Journal of the Korean Society for Information Science and Technology, vol. 11, **(2012)**, pp. 61-74.
[7] M. S. Yeom, "Understanding Consumer Showrooming Behavior Applying Rational Behavior Theory", Distribution Studies, vol. 20, no. 4, **(2015)**, pp. 79-103.
[8] A. S. Oh, "Smart Factory Logistics Management System Using Bluetooth Beacon Based Indoor Location Tracking Technology", The Journal of the Korea Information and Communications Society, vol. 19, **(2015)**.
[9] H. S. Choi and S. B. Yang, "A Study on the Factors Affecting the Intention of Switching from Online Shopping to Webrooming", Intelligent Information Research, vol. 22, no. 1, **(2016)**, pp. 19-41.
[10] H. G. Kim, "Spur building of On/Off convergence 'Omni-channel'... Distribution circles 'Destroy the wall of space/time'", Kookmin Daily News, http://news.kmib.co.kr/article/view.asp?arcid=0922784542&code=11151600&cp=nv, **(2014)**.
[11] Wikipedia encyclopedia, "The definition of Omni-channel", https://en.wikipedia.org/wiki/Omnichannel.
[12] D. S. Hirschberg, "Algorithms for the longest common subsequence problem", Journal of ACM, vol. 24, no. 4, **(1977)**, pp. 664-675.
[13] H. Cao, N. Mamoulis and D. W. Cheung, "Discovery of eriodic patterns in spatiotemporal sequences", IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 4, **(2007)**, pp. 453-467.
[14] J. S. Larson, E. T. Bradlow and P. S. Fader, "An exploratory look at supermarket shopping paths", International Journal of research in Marketing, vol. 22, no. 4, **(2005)**, pp. 395-414.
[15] A. K. Tung, J. Hou and J. Han, "Spatial clustering in the presence of obstacles", Proceedings of the 17th International Conference on IEEE, **(2001)**, pp. 359-367.
[16] V. Srinidhi, "Classification of User Behaviour in Mobile Internet", Asia-pacific Journal of Convergent Research Interchange, HSST, ISSN: 2508-9080, vol. 2, no. 2, **(2016)**, pp. 9-18.
[17] T. S. R. Sowmya, "Cost Minimization for Big Data Processing in Geo-Distributed Data Centres", Asia-pacific Journal of Convergent Research Interchange, HSST, ISSN: 2508-9080, vol. 2, no. 4, **(2016)**, pp. 33-41.

