Journal of Multimedia Information System VOL. 5, NO. 3, September 2018 (pp. 195-200)
http://dx.doi.org/10.9717/JMIS.2018.5.3.195

Analysis of Market Trajectory Data using k-NN

So-Hyun Park[1], Sun-Young Ihm[2], Young-Ho Park[1,*]

Abstract

Recently, as the sensor and big data analysis technology have been developed, there have been a lot of researches that analyze the purchase-related data such as the trajectory information and the stay time. Such purchase-related data is usefully used for the purchase pattern prediction and the purchase time prediction. Because it is difficult to find periodic patterns in large-scale human data, it is necessary to look at actual data sets, find various feature patterns, and then apply a machine learning algorithm appropriate to the pattern and purpose. Although existing papers have been used to analyze data using various machine learning methods, there is a lack of statistical analysis such as finding feature patterns before applying the machine learning algorithm. Therefore, we analyze the purchasing data of Songjeong Maeil Market, which is a data gathering place, and finds some characteristic patterns through statistical data analysis. Based on the results of 1, we derive meaningful conclusions by applying the machine learning algorithm and present future research directions. Through the data analysis, it was confirmed that the number of visits was different according to the regional characteristics around Songjeong Maeil Market, and the distribution of time spent by consumers could be grasped.

Key Words: Data Analytics, Statistical Analytics, K-Nearest Neighbors Algorithm, Point of Sales Data, Trajectory Data.

I. INTRODUCTION

Recently, as the sensor and big data analysis technology have been developed, there have been a lot of researches that analyze the purchase-related data such as the trajectory information and the stay time. Such purchase-related data is usefully used for the purchase pattern prediction [1] and the purchase time prediction [2].

[1] proposed a method of predicting purchase patterns by using Bayesian network, [2] proposed a study to find the periodic pattern in actual product data and use it to predict the purchase point. . [3] conducted research on the purchase pattern extraction using SOM. In order to analyze the trajectory data, [4] performed the performance evaluation of the clustering algorithm for various datasets, [5] proposed an algorithm that can measure the similarity between the traces for analyzing the trajectory data. [6] Classified several kinds of studies related to the analysis of trajectory data and suggested future research directions.

Because it is difficult to find periodic patterns in large-scale human data [2], it is necessary to look at actual data sets, find various feature patterns, and then apply a machine learning algorithm appropriate to the pattern and purpose. Although existing papers have been used to analyze data using various machine learning methods, there is a lack of statistical analysis such as finding feature patterns before applying the machine learning algorithm

Therefore, the contribution of this paper is as follows.

1. Analyzes the purchasing data of Songjeong Maeil Market, which is a data gathering place, and finds some characteristic patterns through statistical data analysis.
2. Based on the results of 1, we derive meaningful conclusions by applying the machine learning algorithm and present future research directions.

The composition of this paper is as follows. Section 2 introduces the related research, and Section 3 explains the process of preprocessing the movement trajectory data of the customers obtained from the sensors installed in Songjeong Meiji market. In Chapter 4, statistical data analysis is performed based on the pre-processed data. In Chapter 5, the experiment is conducted and the results are analyzed. Finally, in Chapter 5, conclusions are made and future research is introduced.

II. RELATED WORKS

Manuscript received July 11, 2018; Revised August 07; Accepted August 13, 2018. (ID No. JMIS-2018-0036)

Corresponding Author (*): Young-Ho Park, Department of IT Engineering, Sookmyung Women's University, Seoul, Republic of Korea, yhpark@sm.ac.kr
[1]Department of IT Engineering, Sookmyung Women's University, Seoul, Republic of Korea, shpark@sm.ac.kr
[2]Department of IT Engineering, Sookmyung Women's University, Seoul, Republic of Korea, sunnyihm@sm.ac.kr
[3]Department of IT Engineering, Sookmyung Women's University, Seoul, Republic of Korea, yhpark@sm.ac.kr

A variety of studies have been conducted to analyze consumer purchase patterns. [1] investigated the know-how of the seller and the consumer's purchase consciousness in order to analyze the consumer's purchase behavior pattern. Based on the surveyed data, they constructed a Bayesian network and proposed a method of predicting purchase patterns. [2] mentions the difficulty of finding periodic patterns in large-scale data generated by humans, and has defined cycle patterns and used them to find patterns in real product data. This study has made research that can be used for marketing for forecasting of purchase time, loss of periodicity, and improvement of customer loyalty. [3] conducted a study of purchasing pattern extraction using SOM based on RFM (Recency, Frequency, Monetary) analysis in ubiquitous commerce.

In addition, various studies have been made on the analysis of the trajectory data. [4] compared the clustering performance of the trajectory data, and the performance evaluation was performed by dividing the som algorithm and the kmeans algorithm into several performance evaluation criteria using various data sets. [5] used a graph to analyze the trajectory data, and proposed a method of adding the MCS (Maximum Common Subgraph) algorithm to the Hausdoff algorithm, which is a widely used distance measurement algorithm, for personalized trajectory data analysis. [6] categorized the studies that have been studied so far for the analysis of the trajectory data into three categories: clustering algorithms, prediction algorithms, and association algorithms. [7] proposed a Taxi-hunting Recommendatio System (Taxi-RS) which provide to both waiting time to take a taxi and location using taxi trajectory data. [8] proposed method which recommend shortest driving route using taxi trajectories.

III. DATA PREPROCESSING

This chapter describes the data collection method, removes data anomalies, and introduces the normalization process using the min-max method.

3.1. Data Collection

Near the Songjeong Station in Gwangju Three sensors were installed at the entrance of Songjeong Maeil Market, incorporating Wifi, BLE and NFC. For Region 1, you can see that it is a dense zone and a station zone as shown in the following figure 1, and Region 2 is a dense area of restaurants. Finally, it can be seen that Region 3 is a dense area of apartment. Region 0 refers to regions excluding regions 1, 2, and 3 in the entire region.

3.2. Data Normalization

Sample data related to purchasing Songjeong Maeil Market are shown in Figure 2. The attribute of the sample data is composed of 4 kinds, and the region attribute means each region region 0, 1, 2, 3 introduced in Section 3.1. Device id means the MAC address of the sensed consumer. First_time_seen refers to the first time in the area, and last_time_seen refers to the last time in the area.

In figure 2, first_time_seen and last_time_seen are the same. In this case, the time to stay is zero, which is when the person is crossing or crossing the area. But that data is 50% of the total data. Therefore, for accurate analysis, data with corresponding values were regarded as missing values and removed.

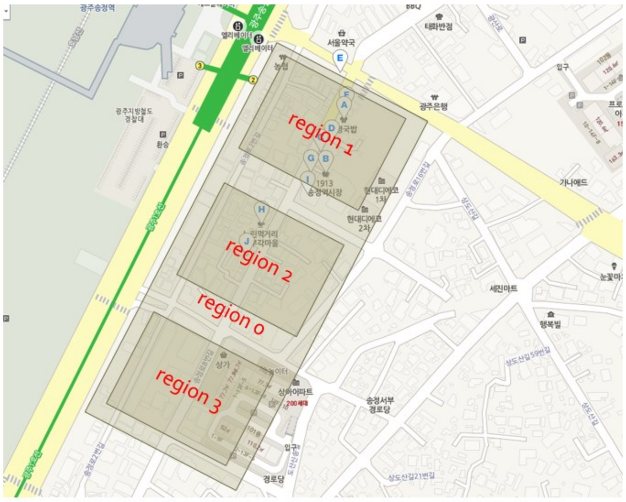
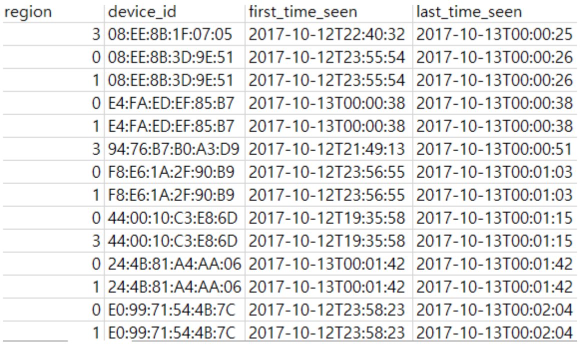


Fig. 1. Integrated module sensor location installed in Songjeong Maeil Market.



region	device_id	first_time_seen	last_time_seen
3	08:EE:8B:1F:07:05	2017-10-12T22:40:32	2017-10-13T00:00:25
0	08:EE:8B:3D:9E:51	2017-10-12T23:55:54	2017-10-13T00:00:26
1	08:EE:8B:3D:9E:51	2017-10-12T23:55:54	2017-10-13T00:00:26
0	E4:FA:ED:EF:85:B7	2017-10-13T00:00:38	2017-10-13T00:00:38
1	E4:FA:ED:EF:85:B7	2017-10-13T00:00:38	2017-10-13T00:00:38
3	94:76:B7:B0:A3:D9	2017-10-12T21:49:13	2017-10-13T00:00:51
0	F8:E6:1A:2F:90:B9	2017-10-12T23:56:55	2017-10-13T00:01:03
1	F8:E6:1A:2F:90:B9	2017-10-12T23:56:55	2017-10-13T00:01:03
0	44:00:10:C3:E8:6D	2017-10-12T19:35:58	2017-10-13T00:01:15
3	44:00:10:C3:E8:6D	2017-10-12T19:35:58	2017-10-13T00:01:15
0	24:4B:81:A4:AA:06	2017-10-13T00:01:42	2017-10-13T00:01:42
1	24:4B:81:A4:AA:06	2017-10-13T00:01:42	2017-10-13T00:01:42
0	E0:99:71:54:4B:7C	2017-10-12T23:58:23	2017-10-13T00:02:04
1	E0:99:71:54:4B:7C	2017-10-12T23:58:23	2017-10-13T00:02:04

Fig. 2. Sample data related to purchasing Songjeong Maeil Market.

In order to apply the machine learning algorithm and get accurate analysis results, data normalization work should be done to make all attribute data fall into similar numerical categories. In this study, all data categories are normalized between 0 and 1 using Min-Max normalization method, which is a typical method of data normalization. The Min-max normalization equation is shown in Figure x.

In order to normalize the i-th data of the attribute x, the difference of $\min(x)$, which means the minimum value of the attribute in the i-th data of the corresponding attribute x, is obtained. It also finds the difference in $\min(x)$ from $\max(x)$, which means the maximum value of the attribute. Divide the two values to obtain the normalized value z^i.

$$z^i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Fig. 2. Min-Max normalization expression

IV. DATA ANALYTICS

In this section, statistical data is analyzed for the number of visits to each area and the distribution of time spent in each area.

In order to analyze the number of visits to each region, the number of visits in each region was counted. The number of visits is shown in Figure 3. The regions 0, 1, 2, and 3 have visited about 630,000, about 210,000, about 220,000, and about 622,000, respectively.

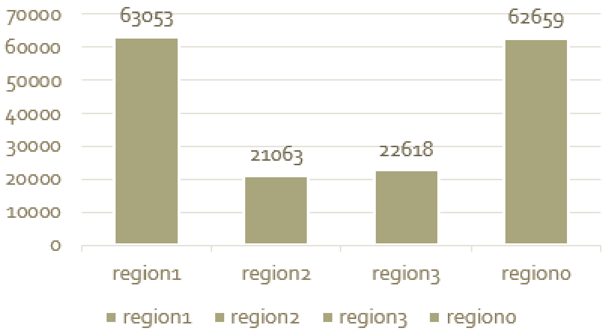


Fig. 3. Visit times in each region.

Region 0 refers to regions excluding regions 1, 2, and 3 in the entire region. The number of visits to region 1 was about 2.9 times higher than that of region 2 and 2.7 times higher than that of region 3. The reason for the difference in the number of visits is that the region 1 is closest to the station and the restaurant is densely packed. Region 2 is the area where restaurants are crowded but is not located in the station area like Region 1. Region 3 is a densely populated area, but there is a big difference in the number of visits because station areas and restaurants are not dense. It is usually a station area, a lot of apartments, and a lot of places where restaurants are crowded. As a result of analyzing the trajectory data obtained from the sensing, the number of visits actually increased in regions having such regional characteristics. Also, the number of visits was found to be much in the order of region 1, region 0, region 3 / region 2.

To analyze the distribution of time spent in each region, we counted the number of visits per time spent in each region. The results of counting the number of visits per stay in each region are as follows. As can be seen in Figure 4, which is a graph showing the distribution of time (within 24 hours) for each region, the staying time is the most, from 0 hours to less than 2 hours, 99%. In addition, as shown in the graph of the distribution chart of the stay time (within 2 hours) of each region, as shown in Figure. 5, 52.79% was the case where the entry time and the leaving time were the same.

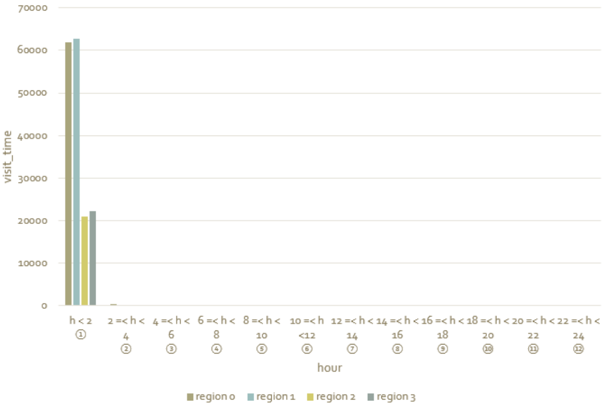


Fig. 4. Time spent in each area (within 24 hours).

Table 1. Number of visits per stay time

Stay time	**Visit time**	**Visit time in stay time / visit time in all stay time**
h < 2 (①)	167804	99
2 =< h < 4 (②)	752	0.4
4 =< h < 6 (③)	265	0.15
10 =< h <12 (⑥)	151	0.09
6 =< h < 8 (④)	131	0.08
8 =< h < 10 (⑤)	154	0.07
12 =< h < 14 (⑦)	74	0.04
22 =< h < 24 (⑫)	16	0.014
14 =< h < 16 (⑧)	10	0.009
20 =< h < 22 (⑪)	0	0.007
16 =< h < 18 (⑨)	12	0.005
18 =< h < 20 (⑩)	24	0

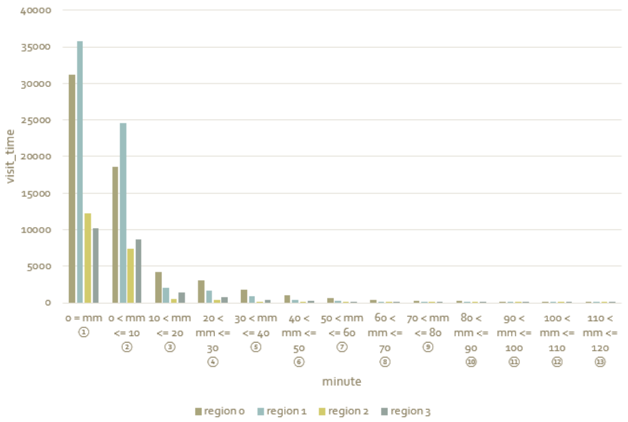


Fig. 5. Time spent in each area (within 2 hours).

Table 2. Number of visits per stay time.

Stay time	Visit time	Visit time in stay time / visit time in all stay time
0 = mm (①)	89429	52.79
0 < mm <= 10 (②)	56217	33.19
10 < mm <= 20 (③)	8148	4.81
20 < mm <= 30 (④)	5875	3.47
30 < mm <= 40 (⑤)	3174	1.87
40 < mm <= 50 (⑥)	1778	1.05
50 < mm <= 60 (⑦)	1045	0.62
60 < mm <= 70 (⑧)	663	0.39
70 < mm <= 80 (⑨)	478	0.28
80 < mm <= 90 (⑩)	335	0.20
90 < mm <= 100 (⑪)	273	0.16
100 < mm <= 110 (⑫)	228	0.13
110 < mm <= 120 (⑬)	161	0.10

V. EXPERIMENT AND RESULTS

In this section, we propose a method of estimating a group of spectators based on visitor information using k-Nearest Neighborhood (k-NN) algorithm. The k-NN is an algorithm that classifies the input data into items belonging to k closest training data in a specific space. Through this, it is possible to provide a service for recommending related products to visitors belonging to the group. The development language is Java language, and the development environment is NetBeans IDE 8.2.

Since there is a proposal in the prediction as the attribute of the present data, the algorithm is operated by adding the virtual attribute data. When creating virtual attribute data, data was generated according to the following rules in Figure 6.

Experimental results show that the amount of data is large and the prediction accuracy is low due to the addition of virtual data attributes. If gender and age data are added, we will proceed with research to predict customer preference through analysis of various machine learning algorithms and to utilize it in consumer market.

VI. CONCLUSION AND FUTURE WORKS

In this study, some characteristic patterns were found through statistical data analysis of data related to purchase of daily market in Songjeong Station. Based on these feature patterns, k-NN algorithm is applied. The result of the prediction accuracy is low due to the addition of virtual data attributes and data due to the lack of attributes in the existing Songjeong Daily Market purchase related data. We analyzed the trajectory data related to consumption in Korea through the data obtained from the sensors installed in Songjeong Maeil Market and hope that it will play an important role in revitalizing the Korean consumer market. Through the data analysis, it was confirmed that the number of visits was different according to the regional characteristics around Songjeong Maeil Market, and the distribution of time spent by consumers could be grasped.

Future research will be conducted to predict the trajectory of visitors based on time series data related algorithms and top-k research.

1. Gwangju, where the Songjeong Maeil Market is located, assumes that the visitor staying far from the area will be short of time (1: Gwangju, 2: Daejeon, 3: Daegu, 4: Busan, 5: Seoul Incheon, 6: Ulsan)
2. Randomly set the visitor's residence area to a value between 1 and 6 if the time to stay is zero. If the time of stay is 0, it is set to a random value because it is mostly a passing vehicle. Therefore, a random setting of 1 to 6, which means all regions
3. If the stay time is more than 0 and less than 60 minutes, set it to 6 (Ulsan), the most remote residential area in Gwangju
4. Set your residential area to 5 (Seoul) if your stay is longer than 60 minutes and less than 120 minutes
5. If the time of stay is more than 120 minutes and less than 180 minutes, it is randomly set to a value between 2-4 in Busan, Daegu, Daejeon.

Fig. 5. Virtual Data Generating Rules.

Gender, Age, Visit Time
Various Datamining Alogirhtm
Customer Preference

Fig. 6. Improved Design Version of Analysing Market Trajectory Data.

ACKNOWLEDGEMENT

This research was financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program.

REFERENCES

[1] J. S. Hwang, S. Y. Pi, C. S. Son, H. M. Chung, "A Purchase Pattern Analysis Using Bayesian Network and Neural Network" , *International Journal of Fuzzy Logic and Intelligent systems*, vol. 15, no. 3, pp. 306-311, 2005

[2] J. W. Kim, W. S. Lee, "Purchase Prediction and Marketing Utilization Through Pseudo Periodic Pattern Analysis", in *Proceedings of Korean Institute of Information Technology Summer Conference*, pp. 52-55, June. 2017.

[3] Y. S. Cho, S. C. Moon, K. H. Ryu, "SOM Clustering Method based on RFM Analysis for Predicting Customer Purchase Pattern in u-Commerce", *Journal of The Korea Society of Computer and Information*, vol. 21, no. 2, pp. 185-187, July. 2013.

[4] N. Y. Kang, J. Y. Kang, H. S. Yong, "Performance Comparison of Clustering Techniques for Spatio-Temporal Data", *Journal of Korea Intelligent Information System Society*, vol. 10, no. 2, pp. 15-37, Nov. 2004.

[5] J. H. Hong, K. S. Park, Y. K. Han, Y. K. Lee "A Method for Measuring Similarity between Trajectory Graph Sets", *Journal of Korea Intelligent Information System Society*, vol. 40, no. 3, pp. 153-158, 2013.

[6] M. Y. Jang, M. Yoon, J. W. Chang, "A Survey on Moving Object Trajectory Mining Techniques in Location-based Services", *Journal of Korea Intelligent Information System Society*, vol. 28. No. 1, pp. 67-68, 2012.

[7] X. Xu, J. Zhou, Y. Liu, Z. Xu, X. Zhao, " Taxi-RS: Taxi Recommendation System Based on Taxi GPS Data", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1716-1727, 2015.

[8] W. Yang, X. Wang, S. M. Rahimi, J. Luo, " Recommending Profitable Taxi Travel Routes Based on Big Taxi Trajectories Data", in *Proceeding of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2015)*, pp. 370-382, 2015.

Authors

So-Hyun Park is a PhD student at Dept. of IT Engineering, Sookmyung Women's University. She received her Bachelor degree from Piano, and Master degree from Multimedia Science both from Sookmyung Women's University. Her interests include multimedia database, deep learning, smart education, etc.

Sun-Young Ihm is a Visiting Professor of Dept. of IT Engineering, Engineering School at Sookmyung Women's University. Her research interests include big data, data analysis, machine learning, top-k query, index building and database.

Young-Ho Park is a Professor of Dept. of IT Engineering, Engineering School at Sookmyung Women's University. Recently, his research interests include data science based-on data analytics, Database Management Systems (DBMS), Information Retrieval (IR), Machine Learning (ML), XML, and IT Convergence with other fields such as music, design, economy, business management, advertisement, bio-informatics, etc. He received his PhD degree in Department of Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in 2005. His PhD research includes efficient query processing in heterogeneous XML documents. He received his BS and MS degrees in Computer Engineering from the Dongguk University in 1990 and 1992, respectively. He had worked for Electronics and Telecommunication Research Institute (ETRI) as a senior research staff at the ISDN Administration & Maintenance Division for TDX-10 ISDN, the Real-Time DBMS Division and the Real-Time Operating System Division from 1993-1999. He had also worked for the Advanced Information Technology Research Center (AITrc), KAIST as a Post-Doctoral Researcher from 2005-2006 after receiving PhD degree.