Spatio-temporal Analysis of Retail Customer Behavior based on Clustering and Sequential Pattern Mining

Hao Chen
Guangxi Tobacco Industrial Co.,Ltd
Nanning, P.R.China, 530001
howard.chenhao@qq.com

Sizhe Yu
Mathematics School, Sichuan University
Chengdu, P.R.China, 610065
yusizhe_1@126.com

Feijie Huang
Guangxi Tobacco Industrial Co.,Ltd
Nanning, P.R.China, 530001
279301345@qq.com

Bing Zhu*
Business School, Sichuan University
Chengdu, P.R.China, 610065
zhubing1866@hotmail.com

Lin Gao
Qingdao Tobacco Co.,Ltd
Qingdao, P.R.China, 266072
29024405@qq.com

Cheng Qian
Business School, Sichuan University
Chengdu, P.R.China, 610065
qiancheng1948@outlook.com

***Abstract*—In the era of big data, the availability of large amounts of spatio-temporal data provides a new path for customer analytics. However, the related research in the retail industry is underdeveloped. In this paper, we introduce spatio-temporal data mining into retail customer analytics and conduct experiments on large real-world data sets from a retail company containing millions of customer purchase records. Spatio-temporal clustering and a new hybrid sequential pattern mining method are used to discover the characteristics of customer behavior at the aggregation as well as the individual level. The typical spatial and temporal distribution of customers and main customer clusters are obtained. Some interesting sequential purchase patterns are also found. Our research will provide not only a new analytic framework for academia but also some guidelines for better development of marketing strategies in the retail industry.**

***Index Terms*—Retail customer analytics, Spatio-temporal data, Spatio-temporal clustering, Sequential Pattern Mining**

I. INTRODUCTION

Customer analytics is an inevitable hot topic in the retail industry [1]. In the context of big data, enterprises can learn the behavior of customers through appropriate modeling methods. For example, how much customers will spend for a certain product or service, when and where different customer groups prefer to shop, and what is the pattern of their purchase records. All these studies can help companies better serve their customers, leading to more profits for businesses. In order to ensure the accuracy of these studies, enterprises need not only using advanced technologies but also different types of data.

Due to the popularization of digital equipment such as sensors and mobile phones, a massive amount of spatio-temporal data is available. Spatio-temporal data mining has become an increasingly important research field in recent years [2]. Spatio-temporal data mining focuses on both time and space dimensions of the study object and can provide more insights into customer behavior from different perspectives. Some research has attempted to applied spatio-temporal data to customer analytics [3][4], but the related research is still in its infancy, especially in the retail industry.

In the context above, we introduce spatio-temporal data mining techniques into retail customer analytics. We try to investigate customer behavior from two perspectives: spatial and temporal pattern of customer purchase behavior at the aggregation level and the characteristics of customer purchase trajectory at the individual level. A large data set with more than 1.4 million customer purchase records from a retail company will be used in the study. Specifically, we first analyze the spatio-temporal distribution of customer behavior in both space and time dimensions, then we perform spatio-temporal clustering to discover the aggregation of hot-spots. A newly proposed spatio-temporal clustering method called ST-DBSCAN [5] is utilized and several representative clusters are obtained. The typical characteristics of each cluster are also studied. Second, we perform sequential pattern mining to explore the potential pattern in the customers' purchase sequence. A hybrid semantic sequential pattern mining technique is used which combines of Prefix-Span algorithm and Point of Interest (POI) semantic analysis [6]. We discover several frequent sequential patterns in customer shopping trajectories. To the best of our knowledge, all the two types of analysis have never been touched in previous research. Our research will not only provide a new analytic framework for academia but also some guidelines for customer relationship management in the retail industry.

The rest content of this paper is mainly divided into four parts. In Section II, we introduce related work on spatio-temporal data mining. In Section III, we briefly introduce the techniques used in the experiments. Section IV shows he

978-1-7281-9740-1/20/$31.00 ©2020 IEEE

experiments settings and analysis of the results. In Section V, we conclude the whole paper and give practical implications as well as future directions of our research.

II. Related Work

Spatio-temporal data mining is becoming increasingly important in data mining for the last decade. Most of the research in spatio-temporal data mining can be roughly divided into two groups: one focuses on the data mining task and the other on applications. The former group pays more attention to develop different techniques to improve the performance in tasks such as clustering, predictive learning, change detection, frequent pattern mining, anomaly detection, and relationship mining [2]. The latter group attempts to apply the corresponding techniques to solve the problem in different fields, such as social sciences, climate science, neuroscience, epidemiology, and transportation [7]. Spatio-temporal data provides new research opportunities to customer analytics due to the increasing availability of a large amount of spatio-temporal data collected by mobile phones and sensors, but there are only a few research papers [3][4], especially in the retail industry.

The spatial-temporal distribution law of different events in an interesting research issue in spatio-temporal data mining. Previous research mainly used the statistics method. For example, Nanbo et al. characterized the spatio-temporal distribution of Ebola virus proteins and RNA during virus replication [8]. Wang et al. analyzed spatio-temporal distribution of lifespan indicators according to population censuses data [9]. Recently, the method of spatio-temporal clustering is developing quickly. The existing methods mainly include spatio-temporal scanning statistical methods [10][11], density-based methods [3][12][13] and space-time distance-based methods [14]. Kisilevich et al. made a systematic review on these methods [15]. Among them, the density-based method ST-DBSCAN proposed by Birant [5] has shown its advantage of noise-tolerance and nonlinear boundary. Spatio-temporal clustering is suitable for mining the concentrated distribution of data, but there are no related applications in customer analytics yet.

Sequential pattern mining is a significant mining task in spatio-temporal data mining. It can help to find frequent sequences of moving objects and has been extensively studied. Some people tried to develop new algorithms to improve the performance of sequential pattern mining. For example, Cao et al. employed a substring tree structure to solve the fuzziness of locations and the identification of non-explicit pattern instances [16]. Other researchers have applied the techniques into many domains such as travel recommendation [17], life pattern understanding [18], and next location prediction [19]. These studies successfully discover the potential patterns of sequences. In the retail industry, Clemente et al. used credit card data to classify sequences of purchases [20], but their method has not considered spatial behavior information.

III. Methodology

In retail customer analytics, there are two important problems to be addressed. First, when and where the customer events gather. Second, what are the frequent customer behavior sequences, especially the shopping location sequences. Addressing these two problems can help managers allocate precious business resources and develop suitable marketing strategies. In this paper, we use the spatio-temporal clustering algorithm called ST-DBSCAN to help solve the first problem. Then a hybrid semantic sequential pattern mining technique based on the Prefix-Span algorithm and POI semantic analysis is used to address the second one.

A. ST-DBSCAN clustering

ST-DBSCAN is an advanced method of spatio-temporal clustering proposed by Birant [5], which uses multi-information to explore the relationship of the points in terms of both time and space. Based on DBSCAN's definition of spatial neighborhood, ST-DBSCAN considers the temporal neighborhood according to the adjacent time points, so as to use both spatial and temporal attributes for clustering decisions.

ST-DBSCAN algorithm requires four parameters $Eps1$, $Eps2$, $MinPts$, and $\Delta\epsilon$. $Eps1$ is the distance parameter for spatial attributes (i.e., latitude and longitude). $Eps2$ is the distance parameter for non-spatial attributes (i.e., temporal attributes). In this research, we apply the Euclidean distance metric as follows:

$$Eps1 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

$$Eps2 = \sqrt{(t_1 - t_2)^2 + (t_1 - t_2)^2} \tag{2}$$

where x_1, y_1 and x_2, y_2 are the latitude and longitude of the object and its neighbour, t_1, t_2 are the values of their time attribute. $MinPts$ is the minimum number of points within the spatio-temporal neighborhood to determine $Eps1$ and $Eps2$. $Minpts$ is usually set to be $\ln(n)$, where n is the size of the data set. In particular, we need to determine $Eps1$ and $Eps2$ according to $Minpts$ and the actual situation. The last parameter $\Delta\epsilon$ is used to prevent the situation of combined clusters because of the little differences in temporal values. ST-DBSCAN has two obvious advantages. The first one is that it can automatically remove the noises. The second is that the boundaries of the clusters are nonlinear, which can avoid some tough problems such as class overlap.

B. Semantic sequential pattern mining

To perform sequential pattern mining, we should first generate the symbol sequences of shopping places. To get the elements in the sequences, we convert the "longitude" and "latitude" attributes to semantic labels representing shopping places based on POI technology. According to the POI classification system of *Amap*, all places are divided into 12 categories. They are Life service, Food service, Shopping and Entertainment, Education, Residence, Company, Traffic, Sports Leisure, Scenery, Government, Medical care, and Others. According to the longitude and latitude, we match each purchase record to the corresponding class through the nearest POI locations of them.

After the transformation, the Prefix-Span algorithm proposed by Han [6] is used to discover the frequent patterns of sequences. The prefix-Span algorithm is based on prefix and projection. Prefix refers to the sub-sequence of the preceding part of a sequence, and projection refers to the complement set corresponding to a prefix of the sequence, also known as the suffix. For example, a prefix of the sequence $< a(ABC)dc(ab) >$ is $< aa >$, and its projection is $< (\_bc)dc(ab) >$. Similar to the Apriori algorithm, Prefix-Span starts from the prefixes in the length of 1 and searches the corresponding projection database to get frequent sequences. Next, it continues to discover the prefixes in the length of 2 recursively. It stops until it can no longer find longer prefixes. Then, frequent semantic patterns of customer behavior are obtained.

IV. Experiments and Results

This part shows the experimental setting and the analysis of the results. In the experiment, we pre-process the raw data and convert it into an easy operated form. After that, we will use the methodology described in section III to analyze the data and explain the results.

A. Experimental Settings

The data set used in the experiments comes from a retail company. It contains over 1.4 million real product purchase records in Q city, collected by scanning on the *WeChat* QR Code. Each record contains the following information: customer ID, item name, longitude, latitude, time, and city district. As a necessary pre-processing step, we deleted the customers who have less than 5 records. After being preprocessed, the data set remains to contain more than 1.2 million records.

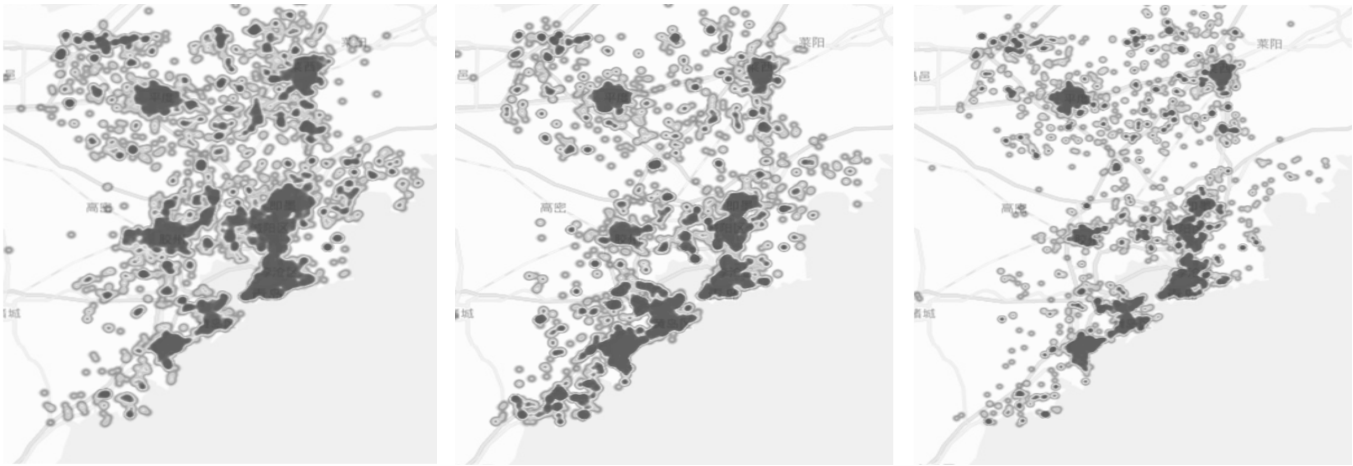
According to the definition of ST-DBSCAN, we should have set $Minpts$ to 11. However, our preliminary experiments showed the sample points were so dense that there were much more than 11 points in most of the neighborhood, but the points on the boundary were relatively sparse. So we reduce $MinPts$ appropriately to neutralize the excessive aggregation of internal points. We finally choose $MinPts = 8$. Then we set $Eps1 = 15000$ and $Eps2 = 30$.

In sequential pattern mining, we only considered the sequence within a week. A big problem of sequential pattern mining is that it only focuses on the sub-sequence, which means the time span has not been considered. For example, the result may discover a sub-sequence $<$ Medical care $\rightarrow$ Sports Leisure$>$, while there is a gap in length of half a year between the two records. To avoid this, we sliced the time span by week. That means each behavioral sequence only includes all the purchase records of one customer in one week.

B. Spatio-temporal clustering

Fig. 1 shows the distribution of purchase records in different time periods of a day. The south part of the map, where the hot spots are concentrated, is the main urban area of Q city, while the northern parts are surrounding cities and counties. We can see from Fig. 1 that in the afternoon (18 p.m.) the purchase frequency is higher, and the purchase is more intensive in the city. While at night (22:00 p.m.), the number of purchase records of cities decreases significantly, but that of surrounding cities and counties increases. This phenomenon shows that part of the population of cities and counties around Q city will enter the urban area in the daytime.

Fig. 2 shows the results of ST-DBSCAN clustering in three-dimensional space, in which four main clusters were obtained. The points of each cluster are respectively assigned with different gray scales. We can see from Fig. 2 that these clusters have very different distributions with regards to time and space. To show the results more clearly, Table I summarizes the statistical properties of different clusters and Fig. 3 depicts the spatial distribution of the clusters. From the clustering results in Table I and Fig. 3, we can get a lot of valuable



(a) All day (b) 18 p.m. (c) 22 p.m.

Fig. 1. Distributions of Purchase Records in Different Time Periods

information for the retail company. Cluster1 and Cluster2 show that people are more likely to buy a product before and after work near their companies. Cluster3 and Cluster4 indicate that the working and rest time of surrounding counties is later than that of cities. The pace of life will be slower, and the product purchase density is even higher than that of cities. Retail companies can further analyze the purchase hot-points of different commercial areas at different time and promote the intercity allocation of products and marketing resources reasonably. intercity

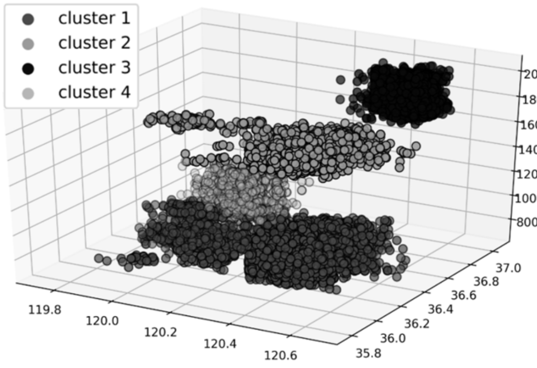


Fig. 2. Spatio-temporal Clustering Result

C. Frequent Customer Sequential Pattern

By using the POI semantic analysis, we discover a phenomenon that most of the sequential patterns contain only one sort of POI class. For example, there are 75,388 single-valued sequences <Shopping & Entertainment → Shopping & Entertainment → Shopping & Entertainment> and 69623 <Life service → Life service → Life service>. It shows that most customers prefer to buy products in places with the same type, even in a fixed location.

To study more on the transition between different POI classes, we combine the continuous elements of the same class into one element to ensure that the nearby elements are different. After that, we discover two similar frequent patterns by using the hybrid sequential pattern mining algorithm as follows:

- Pattern 1 : <Life service → Shopping & Entertainment → Life service>
- Pattern 2 : <Shopping & Entertainment → Life service → Shopping & Entertainment>

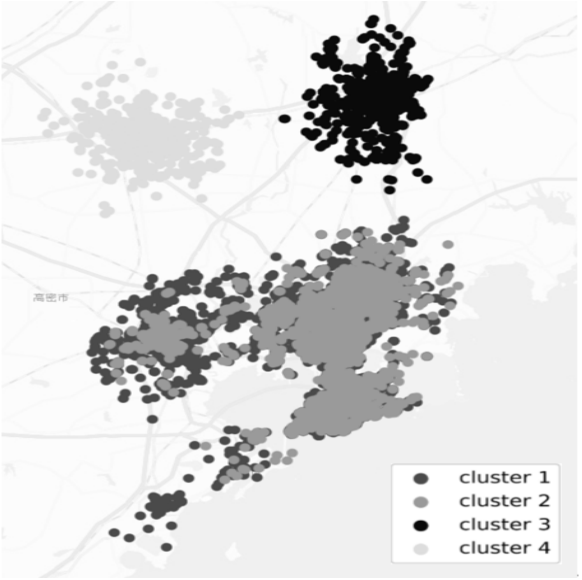


Fig. 3. Spatial Distribution of Different Clusters

These two patterns can be assorted to one category, showing that customers are more likely to switch between places of “Shopping & Entertainment” and places of “Life service”. This sequential pattern is visualized in Fig. 4.

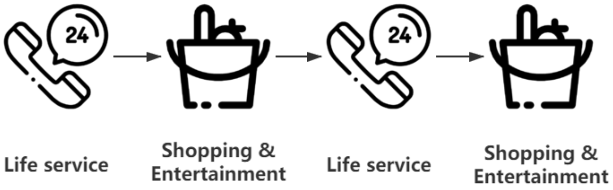


Fig. 4. Most Frequent Purchase Pattern

Beside the two patterns, We also find two less frequent patterns as follows:

- Pattern 3 : <Shopping & Entertainment → Food service → Shopping & Entertainment>
- Pattern 4 : <Shopping & Entertainment → Traffic → Shopping & Entertainment>

All these results can help not only discover the lifestyle of customers but also predict their purchase trajectories. Moreover, we also analyzed the difference between frequent patterns in different seasons. The result shows that although there

TABLE I
STATISTICAL PROPERTIES OF DIFFERENT CLUSTERS

Cluster	Quantity	Mean of Time	Maximum of Time	Minimum of Time	Mean of location
1	97706	8:44	7:04	10:34	(120.340, 36.281)
2	45182	16:50	15:51	17:42	(120.340, 36.279)
3	42289	18:33	17:00	20:14	(120.523, 36.869)
4	41539	8:50	7:13	10:32	(119.971, 36.782)

is hardly variation, the frequency of Pattern 2 (<Shopping & Entertainment → Life service → Shopping & Entertainment>) decreases in winter.

V. Conclusion

The emergence of spatio-temporal data analysis technology provides a great opportunity for retail customer research. In this paper, we apply spatio-temporal data mining into the retail industry. Based on the characteristics of spatio-temporal data, we analyze the spatio-temporal distribution and sequential patterns of customers' purchasing behavior. Specifically, we discover four distinct clusters by using the spatio-temporal clustering, which reflects the aggregation of customer purchase events. In customer sequential pattern mining, we discover some common sequence patterns of shopping places. The results can help the enterprise enhance the understanding of their customers and better allocate business resources at the proper time and places, which will thus provide customers with better services.

Further work will mainly focus on the two directions. First, we will adjust the granularity of spatial clustering analysis to associate the clusters with different districts. Second, we will analyze the similarity and differences of the behavior sequences among different customer groups.

References

[1] Germann F, Lilien G L, Fiedler L, et al. Do retailers benefit from deploying customer analytics? Journal of Retailing, 2014, 90(4): 587-593.

[2] Atluri G, Karpatne A, Kumar V. Spatio-temporal data mining: A survey of problems and methods. ACM Computing Surveys (CSUR), 2018, 51(4): 1-41.

[3] Urkup C, Bozkaya B, Salman F. Customer mobility signatures and financial indicators as predictors in product recommendation. PloS one, 2018, 13(7), e0201197.

[4] Kaya E, Dong X, Suhara Y, Balcisoy S, Bozkaya B. Behavioral attributes and financial churn prediction. EPJ Data Science, 2018, 7(1): 41.

[5] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data & Knowledge Engineering, 2007, 60: 208–221.

[6] Han J, Pei J, Mortazavi-Asl B, et al. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth, Proceedings of the 17th international conference on data engineering, IEEE, Washington, DC, USA, 2001: 215-224.

[7] Wang S, Cao J, Yu P. Deep learning for spatio-temporal data mining: A survey, arXiv:1906.04928, 2019.

[8] Nanbo A, Watanabe S, Halfmann P, et al. The spatio-temporal distribution dynamics of Ebola virus proteins and RNA in infected cells. Scientific Reports, 2013, 3: 1206.

[9] Wang S, Luo K, Liu Y. Spatio-temporal distribution of human lifespan in China. Scientific Reports, 2015, 5: 13844.

[10] Gaudart J, Poudiougou B, Dicko A, et al. Space-time clustering of childhood malaria at the household level: a dynamic cohort in a Mali village. BMC Public Health, 2006, 6: 1–13.

[11] Kulldorff M, Heffernan R, Hartman J, et al. A space-time permutation scan statistics for disease outbreak detection. PLoS Medicine, 2005, 2: 216–224.

[12] Wang M, Wang A, Li A. Mining spatial-temporal clusters from Geodatabase. Proceedings of International Conference on Advanced Data Mining and Applications, 2006, 4093: 263–270.

[13] Pei T, Zhou C, Zhu A, et al. Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise. International Journal of Geographical Information Science, 2010, 24: 925–948

[14] Zaliapin I, Gabrielov A, Keilis-borok V, et al. Clustering analysis of seismicity and aftershock identification. Physical Review Letters, 2008, 018501: 1–4.

[15] Kisilevich S, Mansmann F, Nanni M, et al. Spatio-temporal clustering. In: Data Mining and Knowledge Discovery Handbook, New York: Springer Press, 2010: 855–874.

[16] Cao H, Mamoulis N, Cheung D W. Mining frequent spatio-temporal sequential patterns. Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05), IEEE, 2005: 82–89.

[17] Zheng Y, Zhang L, Ma Z, Xie X, Ma W. Recommending friends and locations based on individual location history. ACM Transactions on the Web (TWEB), 2011, 5(1): 5-44.

[18] Ye Y, Zheng Y, Chen, Feng J, Xie X. Mining individual life pattern based on location history. Proceedings of the 10th IEEE International Conference on Mobile Data Management IEEE, 2009: 1-10.

[19] A. Monreale, F. Pinelli, R. Trasarti, F. Giannotti. WhereNext: A location predictor on trajectory pattern mining. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009: 637-646.

[20] Di Clemente R, Luengo-Oroz M, Travizano M, et al. Sequences of purchases in credit card data reveal lifestyles in urban populations. Nature Communications, 2018, 9(1): 1-8.