Mining Location-based Service Data for Feature Construction in Retail Store Recommendation

Tsung-Yi Chen[1], Lyu-Cian Chen[2], Yuh-Min Chen[3]

[1]Department of Electronic Commerce Management and Department of Information Management, Nanhua University, Chiayi County, Taiwan (R.O.C.)
tsungyi@mail.nhu.edu.tw
[2]Institute of Manufacturing Information and Systems, National Cheng Kung University, Tainan City, Taiwan (R.O.C.)
lcchen1993@gmail.com
[3]Institute of Manufacturing Information and Systems, National Cheng Kung University, Tainan City, Taiwan (R.O.C.)
ymchen@mail.ncku.edu.tw

Abstract. In recent years, with the popularization of mobile network, the location-based service (LBS) has made great strides, becoming an efficient marketing instrument for enterprises. For the retail business, good selections of store and appropriate marketing techniques are critical to increasing the profit. However, it is difficult to select the retail store because there are numerous considerations and the analysis was short of metadata in the past. Therefore, this study uses LBS, and provides a recommendation method for retail store selection by analyzing the relationship between the user track and point-of-interest (POI).

This study uses regional relevance analysis and human mobility construction to establish the feature values of retail store recommendation. This study proposes (1) architecture of the data model available for retail store recommendation by influential layers of LBS; (2) System-based solution for recommendation of retail stores, adopts the influential factors with specified data in LBS and filtered by industrial types; (3) Industry density, area categories and region/industry clustering methods of POIs. Uses KDE and KMeans to calculate the effect of regional functionality on the retail store selection, similarity is used to calculate the industry category relation, and consumption capacity is considered to state saturation feature.

Keywords: Urban mining, Spatial and temporal data mining, Location-based Service, Retail store recommendation

1 Introduction

Based on the well-developed LBS, many data are available for user behavior analysis. For example, the POI recommendation system uses the user track in location-based social network (LBSN) to analyze the user preference for recommendation. The POI recommendation system has been discussed extensively in previous studies [1-2], most

P. Perner (Ed.): ICDM 2017, LNAI 10357, pp. 68–77, 2017.
DOI: 10.1007/978-3-319-62701-4_6

of which concerning the influential factors. The adaptive or personalized recommendation system uses data mining to obtain the users' personal data, so as to estimate their behaviors, characters and values [3].

These online user behavior evaluation methods based on internet mining are tolerable analysis materials for enterprises. In order to gain commercial profit, the customer segments and purchasing behavior are obtained by user behavior analysis. There are diversified analysis tools and purposes, the decision of retail location plays an important role in the purposes [4]. The data analysis technique derived from advanced technology takes a share gradually. The analysis of customer segments provides effective assistance for enterprises in deciding retail store, the complicated market survey is no longer a good scheme. Using big data to analyze the relationship of consumers with retail stores or other competitors by cluster, gravity model and neural networks has become an important reference for large enterprises to expand retail stores [5].

However, the traditional cluster analysis and gravity model methods rely on extensive calculation and data, while requiring the use of geography information system (GIS) [5]. The data collection process is very difficult, and the technical threshold is relatively high. As the present data is rapidly and frequently recorded via internet, this study adopts LBS to solve the problem of insufficient data sources in the decision-making of retail store by the location information and implicit user behavior.

Since 2010, there are many studies on the LBSN data analysis, most of which analyzed the relationship between mass user data and POIs according to the check-in data of users, so as to know the implicit information of preferences, sequential behaviors and social circle generated when the user checks in [6-7]. In addition, a few studies explored the relationship between locations and the popularity in LBS, and used machine learning to predict the possible optimum retail store location [4].

Most studies concerning LBSN intended to analyze the user behavior, while considering the factors of community, individual preference, regional composition and geographic property, but not the influence of all factors on the decision-making of retail store [1-3, 7, 9]. Therefore, this study discusses the data properties of LBS, point of view of marketing and the relationship inside the regional relationship of retail store and the human mobility, in order to develop a recommendation model for retail store expansion.

To sum up, although with well-developed LBSN and convenient techniques of data collection, there still lacks a recommendation method for retail stores based on LBS data mining. Therefore, this study develops the following purposes:

1. To build an influential factors model for retail stores recommendation system by POI recommendation, so as to identify the data types beneficial to retail store recommendation in LBSN.
2. To establish the retail store recommendation system architecture with LBS data as source.
3. To establish reasonable corresponding recommendation features according to the established influential factors by point of view of marketing.

2 Related work

In the LBSN, the sequential series connection of user's location behavior is called sequential behavior. A lot of useful information can be explored from this behavior, which may contain the information of user's daily routines, preferences and life circle. The sequential behavior is often discussed together with users' social network, that is social influence study. The social network of target user is brought into analysis to obtain the relationships of following between users and their friends. Moreover, multiple sequential behaviors are connected in series to estimate the diffusion relationship by algorithm [1]. Yang et al. proposed how to establish the users' activity characteristics in geo-spatial, so as to reduce the complexity in multi-dimensional condition [8].

The POIs, also known as venues, refers to the stores or attractions in the LBS graphical information. The POI recommendation system is a very important part of the urban mining research. The unique influence model of POI recommendation system is based on the factors in LBSN, such as social factors, individual preference, popularity of POI and geographic position. The target groups with close properties are identified by collaborative filtering, and the probable mobility is predicted for recommendation [3]. Ye also used collaborative filtering technique to calculate the similarity between social influence and geographical influence, and combined two influence aspects to build a fusion framework. With the user preferences for ranking, to achieve the best POI recommendation accuracy [9].

The factors that may be considered in POI recommendation model include social influence, individual preference, popularity of POI and geographic position. The individual preference uses classification of POI properties as main basis, the historical records of recommended target in LBS are used as recommendation reference for the recommended target. The popularity of POIs represents the recommendation value of the locations by the human clustering and following suit that is an important basis of recommendation. The geographic position also has significant effect on the recommended target. If it is too far from the user's relative position, or outside the life circle, this recommendation would not make sense [1-2].

In urban mining, the human mobility must be caught via internet for crowd monitoring. The best data collection method is to let the target feedback position via GPS in time, so as to implement real-time monitoring. The GPS monitoring has many advantages, such as real-time feedback of the location information, and the coherence for data gathering. Such mass continuous data can be used for crowd monitoring in large-scale activities to avoid dangers [10].

Considering the privacy problem, most studies based on crowd mobility do not use GPS as data source, while some use location-based network for research. Despite of the limited public data, the crowd behavior in location-based network still provide diversified research directions. Yuan et al. used Latent Dirichlet Allocation (LDA) to guess the division and function of regions in city according to the distribution of POIs and human mobility pattern, and presented the information that may be difficult to explore in an automated approach [11].

Developing the human mobility model is a challenge due to the computation complexity. Without any corresponding method to reduce the complexity, it is difficult to obtain the result. Wei et al. proposed the concept of grid that solved this. The adjacent POIs is classified into the same grid, and every frequent trajectory is displayed in region. This method greatly reduces the time complexity for calculating the frequent mobility route [12].

When planning for retail store, enterprises adopt data analysis methods of regression analysis, clustering analysis and gravity models. GIS is also used to identify the segmentation of target customer. However, even with technological assistance, decision-making needs to consider past experiences and common sense in commercial activity. If experience is considered in the decision-making process, the decision will be more accurate [5].

There are numerous influential factors in determining an appropriate location of the store. For the enterprises, the decision-making process is like looking for an appropriate portfolio, determining a correct site can bring enormous profit. The decision factors of retail store can be divided into two categories. One is consumer demand, the other one is appropriate location position. Huff proposed considering customer requirement appropriately, to predict if there was adequate demand at the location for sales estimation [13].

Ghosh & Craig indicated that the factors to be considered in retail store decision-making include the attractiveness of location for consumers. In addition, the competitive relation of the location in the region, the probable preference of customers and the demand for the industry should be considered [14]. Exploring user preference and behavior in urban mining is conducive to business operation. Upon LBS, the consumers' habits can be known by data mining. The analyses of visitors flowing rate and competitors distribution are greatly helpful to commercial activity [15].

3 Influential Factors and System Design

This study plans to design a retail store recommendation system based on LBSN. First, relational concept is imported based on the POI recommendation system in LBS to establish the influential factors related to retail store. The relationship between human mobility and the profit of related industry is observed in the POI recommendation system. The user preference is closely related to the types of industry, the recommended crowd is the target customers. Thus, the retail store recommendation system is built by using POI influential factors for analysis.

3.1 Construct Influential Factors with POI Recommendation System

Most of previous studies on POI recommendation use location influence and social influence to divide the influential factors in POI recommendation, and then use fusion rule, such as sum rule and product rule [1-2]. In addition, there are three major types based on feature extraction [3]. This study adopts in a similar concept, the following three types are described as follows:

- Individual preference: the user's personal preference ratio is defined according to the POI types the target user visited. Or the user's activity sequence is established by spatial temporal activity in the form of time correlation, and then the activity related preference types is established [8].
- Social influence: the social influence considers the user's friends, evaluates if there is stronger influence among the friends. Or disregards the intensity of influence, only aims at the behavioral aspects of friends, and measures the influence model (mobility or behavior sequence) for recommended target.
- Location influence: the most decisive factor in the POI recommendation system is the effect of geographic location. In the measurement of location factor, the location is usually represented by coordinates on a two-dimensional plane. The recommendations are ranked by the distance relation between users and stores. However, the users' individual preference shall be considered first when considering related candidate stores, so as to reduce the computational complexity preliminarily.

To sum up, in the analysis of social influence or individual preference, the influential factors may be correlated with geographic position. Based on the three major influential factors, corresponding to the influence layers of retail store recommendation, relevant information extraction model is built. Three data hierarchies are established by the original LBS data, including primitive layer, mobility and social layer, location layer (Fig.1.).

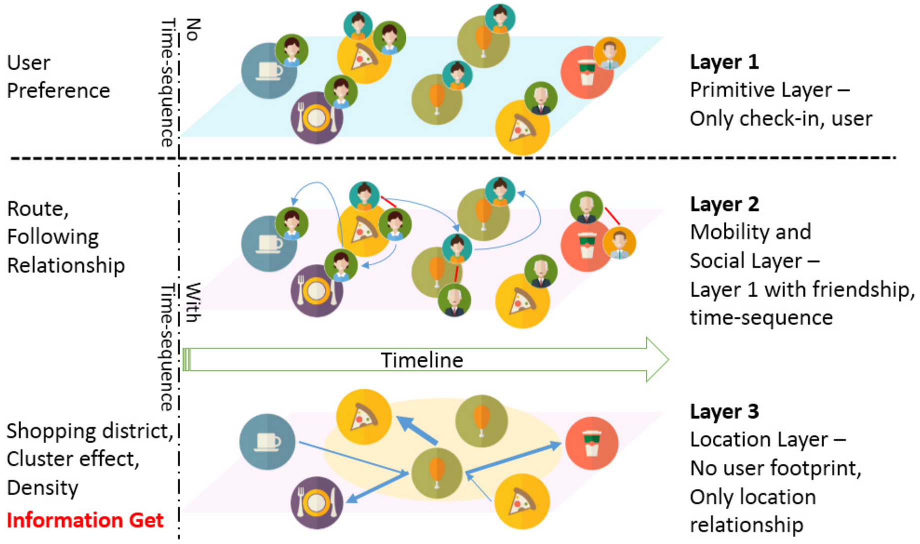

Fig. 1. Information layers for location-based social network

As shown in Fig. 1, the information of Layer 1 can be obtained easily, but the information of Layer 2 and Layer 3 can be obtained after extensive analysis. Moreover, in the study of social influence, it is difficult to obtain social network under the privacy policy. Therefore, removing the effect of social criteria, the features which influence the retail store can be simply divided into two types, one is regional relevance (RR) features, the other one is human mobility (HM) features.

3.2 System Development and Design

In order to recommend appropriate retail store, the enterprise can use the classification mechanism provided by real estate websites for screening. The recommendation system analyzes the implicit influential factors in the filtered candidate store, especially the customer segment and geographic relationships that the real estate websites are difficult to provide. Afterwards, other POIs within a radius of 200 meters from the candidate store and check-in data are collected to analyze the human mobility and industrial relations [4] (Fig. 2.).

As the check-in data is enormous, a filtering mechanism must be established to select the type of industry correlated with enterprise. This mechanism can be relevant information provided by enterprise, or the information derived from web text mining. The more correlated dataset can be filtered out before calculation by the industrial information provided by enterprise and the user preference in LBS, so as to increase the effectiveness of feature values.

Afterwards, the check-in data around each candidate retail store are analyzed, and the computing method of feature value extraction is proposed. The HM features are extracted from the mobility route formed of check-in sequence. In order to establish the mobility route, this study uses the route construction method proposed by Wei [12]. The computational complexity is reduced by grid division, and the industry filtering mechanism is used to select appropriate types to enhance the effectiveness; On the other hand, the RR features are extracted, considering the complex regional relation when the enterprise is selecting retail store location, probably including intensity, saturation, clustering degree, etc.

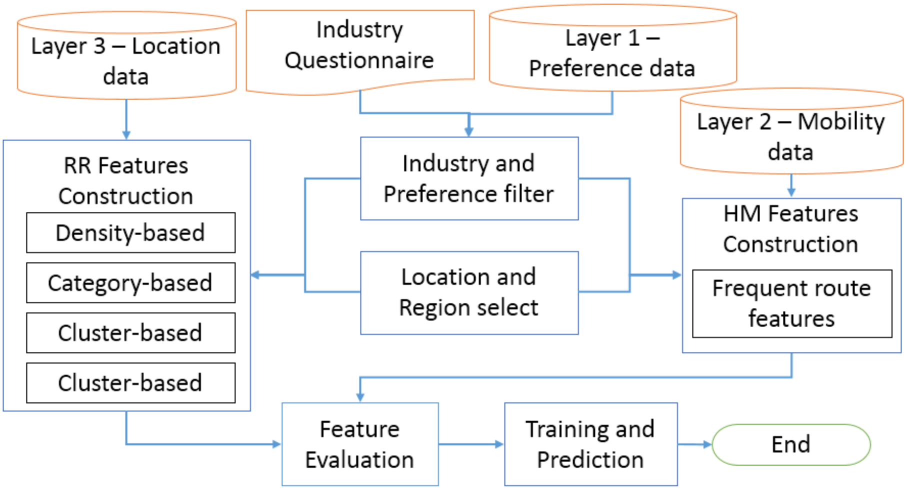

Fig. 2. Framework for feature discovering in retail store recommendation

4 Regional Relevance Features Construction

First we develop features with regional relevance. This paper provide four influence factors for regional relevance, and integrate these methods with industry-preference filter:

Density. The main concern of retailers is if there are competitive stores and crowds nearby the retail store, which are usually reflected in the store rent. The rent at busy areas is extremely high, thus the renters often doubt about the sales revenue and profit. These concerns are often tightly related to the store density. In order to measure the store density in the region, the Kernel Density Estimation (KDE) is used for calculating the density distribution. The KDE is a nonparametric entropy estimation. If the quantity of stores in the region is considered only, there may be fatal loss of some key factors. Only if the density nearby candidate point is calculated, the kernel of trading area is approached in deed. The KDE value of location l is calculated by using two-dimensional KDE, defined as follows:

$$\lambda(l) = \sum \frac{1}{|V_{l,r}|h^2} K(\frac{d_{v,l}}{h}) \text{ , } v \in V_{l,r} \tag{1}$$

Where $d_{v,l}$ represents the distance between POI l and other POIs v in the range, $v \in V_{l,r}$, $V_{l,r}$ represents the set of all POIs in radius r of candidate point l, h is the bandwidth value for KDE calculation. Scott rule is used for estimation, K is kernel function, Gaussian function is used as kernel function in our experiment [11].With density distribution generated by calculating KDE value may not be strong enough, so the industry impact filter is added to KDE function, the POI of different types of industry is given different weights [16]. $W = \{w_c \mid c \in C\}$, represents industry impact quality (IIQ), which determined by enterprise or text mining. C is the set of all POI types, including $c_1, c_2, \dots, c_{|C|}$. Afterwards, the industry weights can be added in to rewrite Eq. (1):

$$\lambda(l) = \sum \frac{1}{|V_{l,r}|h^2} w(v) K(\frac{d_{v,l}}{h}) \text{ , } v \in V_{l,r} \tag{2}$$

Where $w(v)$ represents the weights of each POI v.

Category. In order to measure if the candidate location is in an appropriate region, the industry types distribution in the region must be considered. For example, the food retail shall should be in the residential area or commercial area; bookstores shall be located in the educational area. The industry type distribution in target area l_r is represented by $I_{l,r} = \{i_c \mid c \in C\}$, where i_c represents the vector value of industry type c. In order to check if the recommended region hits our target industry, the similarity between IIQ value W and distribution $I_{l,r}$ is calculated, represented by Euclidean distance φ:

$$\varphi(\mathrm{W}, \mathrm{I}_{l,r}) = \sqrt{\sum_{x=1}^{|C|}(w_x - i_x)^2} \tag{3}$$

Clustering. Clustering is also very important for the store position. The distance between the candidate location and the business district will directly affect the customer's willingness to visit. Therefore, we use the clustering method to select the business district within the region, and after the clustering, analyze the cluster's industrial quality. First, we cluster the POI $V_{l,r}$ in the candidate region l_r by the K-Means clustering algorithm, and get the cluster label for each location. Second, the geometric center of each cluster is obtained, and the distance between the center and the candidate point is calculated. Finally, divide the distance by the industry correlation score φ(Eq. (3)), and then accumulate the results of all the clusters to get the feature score (Algorithm 1).

```
Algorithm 1
Cluster_feature($l$)
    Initialize $V_{l,r}$ and feature_score=0
    clusters=KMeans($V_{l,r}$, $\varepsilon$, MinPts)
    for each cluster in clusters do
        avg = average coordinates of cluster
        $\varphi$ = $\varphi(\mathrm{W}, \mathrm{I}_{cluster})$ according Equation (3)
        accumulate feature_score by $\varphi$/dis-
tance(avg , $l.coordinate$)
    end for
return feature_score
```

Saturation. Market saturation is a very important concept in economics, and its actual degree of saturation depends on the purchasing power of consumers [17]. We can simply denote the saturation degree $\theta_{l,r}$ of a target region l_r as:

$$\theta_{l,r} = -\left|A_{l,r}\right| / \left|V_{l,r}\right| \tag{4}$$

Where $A_{l,r}$ denotes all check-in activity in region l_r, which represents the purchasing power of consumers in this region. This purchasing power divided by the number of businesses in the region. And a negative number indicates that, the greater the number is, the stronger the negative impact on the recommendation. However, this does not fully explain the saturation degree of a single industrial category in the region. In addition, we do not define the saturation point, but only provide a relative comparison. Therefore, we modify the weight distribution of $\mathrm{I}_{l,r}$ to be:

$$\Psi(I_{l,r}) = \begin{cases} 1, i_c \geq \mathrm{m} \\ 0, i_c < \mathrm{m} \end{cases}, c \in C \tag{5}$$

Where m is the relative limit decided by category set C, 0<m<1. And rewrite Eq. (4) based on the concept of industrial distribution:

$$\theta_{l,r,\Psi} = -\frac{|A_{l,r,\Psi}|/|V_{l,r.\Psi}|}{\theta_{max}} \tag{6}$$

θ_{max} represents the max value of saturation degrees for a specified region in the dataset.

5 Conclusion

In order to establish a suitable way for enterprises to find retail stores, we start with LBS and provide a model for mastering geography relation and crowd behavior. There are many implicit relationships that are not easily found through the geo-network. And we use information layered approach, the obvious expression of which can be obtained by data mining technology. In the research, an analytical framework which can take the LBS as the input is designed as the target of system design. In the future, we will use statistical tests to filter the selected features and use machine learning to analyze the best retail locations in the system.

In the selection of features, we take the retailer's point of view, carefully assess the influential factors when choosing retail stores. And then quantify these considerations into the feature values we need, e.g., density distribution, similarity, clustering are all retailers need to consider in detail the characteristics of the region. Based on these analysis of LBS, enterprises can save a considerable amount of cost in data collection. Not only more technical use of data mining methods, but also can improve the accuracy of the forecast.

Reference

1. Zhang, J. D., Chow, C. Y., & Li, Y. (2014, November). LORE: exploiting sequential influence for location recommendations. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 103-112). ACM.
2. Zhang, J. D., & Chow, C. Y. (2015). CoRe: Exploiting the personalized influence of two-dimensional geographic coordinates for location recommendations. *Information Sciences*, *293*, 163-181.
3. Ying, J. J. C., Lu, E. H. C., Kuo, W. N., & Tseng, V. S. (2012, August). Urban point-of-interest recommendation by mining user check-in behaviors. In*Proceedings of the ACM SIGKDD International Workshop on Urban Computing*(pp. 63-70). ACM.
4. Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., & Mascolo, C. (2013, August). Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 793-801). ACM.
5. Hernandez, T., & Bennison, D. (2000). The art and science of retail location decisions. *International Journal of Retail & Distribution Management*, *28*(8), 357-367.
6. Li, Y. M., Chou, C. L., & Lin, L. F. (2014). A social recommender mechanism for location-based group commerce. *Information Sciences*, *274*, 125-142.

7. Wen, Y. T., Lei, P. R., Peng, W. C., & Zhou, X. F. (2014, December). Exploring social influence on location-based social networks. In *2014 IEEE International Conference on Data Mining* (pp. 1043-1048). IEEE.
8. Yang, D., Zhang, D., Zheng, V. W., & Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *45*(1), 129-142.
9. Ye, M., Yin, P., Lee, W. C., & Lee, D. L. (2011, July). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 325-334). ACM.
10. Blanke, U., Tröster, G., Franke, T., & Lukowicz, P. (2014, April). Capturing crowd dynamics at large scale events using participatory gps-localization. In*Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2014 IEEE Ninth International Conference on* (pp. 1-7). IEEE.
11. Yuan, J., Zheng, Y., & Xie, X. (2012, August). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 186-194). ACM.
12. Wei, L. Y., Zheng, Y., & Peng, W. C. (2012, August). Constructing popular routes from uncertain trajectories. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 195-203). ACM.
13. Huff, D. L. (1966). A programmed solution for approximating an optimum retail location. *Land Economics*, *42*(3), 293-303.
14. Ghosh, A., & Craig, C. S. (1983). Formulating retail location strategy in a changing environment. *The Journal of Marketing*, 56-68.
15. Wang, L., Gopal, R., Shankar, R., & Pancras, J. (2015). On the brink: Predicting business failure with mobile location-based checkins. *Decision Support Systems*, *76*, 3-13.
16. Wang, B., & Wang, X. (2007). Bandwidth selection for weighted kernel density estimation. *arXiv preprint arXiv:0709.1616*.
17. O'Kelly, M. (2001). Retail market share and saturation. *Journal of Retailing and Consumer Services*, *8*(1), 37-45.