Prediction on Hot Places in Seoul based on Logistic Regression Analysis

Hye Rim Lee
Dept. of Industrial Engineering
Hongik University
Seoul, South Korea
tky9873@naver.com

Ju Hee Kim
Dept. of Computer Engineering
Hongik University
Seoul, South Korea
joohee1026@naver.com

Hyun Seung Hong
Dept. of Computer Engineering
Hongik University
Seoul, South Korea
harryhong100@naver.com

***Abstract*—**In this paper, web crawling for selecting hot places , collection of data for making analysis/prediction of hot places will take place . Furthermore, visualization of density on each districts of Seoul according to the "number of tags on Instagram" and "number of Starbucks/Mac Donald/Mara soup located" helped us specify certain districts. And by making our own way of scoring system we located certain districts as "hot places". While referring "Seoul Data Plaza" and "Public data portal", information needed for making predictions of hot places was collected. And by using these information collected, "hot place predicting algorithm" was developed, which will be further explained in this paper. Through this predicting algorithm we could confirm 10 future hot places in Seoul.

Keywords—hot place, Seoul data portal, logistic regression, web crawling

I. INTRODUCTION

Currently, big data-based research in the world is being conducted in various fields. In addition, in the age of big data, a myriad of data are pouring out, but most of the data is not utilized. This paper collects various data and predicts hot places in Seoul based on district(Hang-Jeong-Dong). In order to select an existing hot places, there has been a search for Instagram tags (jmt, waiting, date, restaurants, cafe) related to hot places. Moreover, location of the corresponding post was crawled. And, by crawling the location of each stores that are considered as the indicator of hot places (Starbucks, McDonald's, and Mara soup), number of stores located in each district was counted, and moreover, selected the districts that was highly concentrated. In addition, we collected information that could be an indicator among local information published in Seoul Open Data Plaza and public data portal. Among the collected variables, only variables with high correlation coefficients were extracted and used as independent variables when using logistic model. In this paper, the dataset is analyzed and predicted through Logistic Regression. This study would be able to solve the problem of gentrification that is taking place in so called "hot places" located in Seoul, and by providing location insights to future entrepreneurs in their twenties and thirties, it is expected to improve their startup problems. Furthermore, it will be possible to extract the elements found as a feature of the hot place to help revitalize the district.

II. SUBJECT SELECTION, DATA COLLECTION AND DATA PREPROCESSING

Recently, with SNS, people are posting where they visited. In particular, in the case of Instagram, the activation of 20s and 30s, the main consumers of economic activity and mainstream culture, is prominent. Based on this, we will select a hot place and analyze the population/industry status/commercial area/economic characteristics of the district to predict the district that will emerge as a hot place in the future.

Definition of Hot Place

This paper defines hot place as "*A place where the 20 30 population, which has the greatest impact on the consumption trend, has a great favor to visit, making it the center of the trend, and a place that is also recalled a lot in SNS*"

A. Data collection and Data preprocessing

(1) Data web crawling for hot place selection

By searching for tags related to hot places (jmt, waiting, dating, restaurants, cages), we crawled the author, location, creation time, and contents of about 300~400 posts per tag. After that, the collected location was searched by the crawler on Kakao map, and only the location in Seoul was extracted and classified.

The number of stores in each region in Seoul was extracted through a crawler that collects locations after searching for keywords in Starbucks, McDonald's, and Mara soup on the 'Kakao map'. The Python module used to build the crawler is:

- css-selectors
- selenium-webdriver, -chromedriver
- openpyxl

(2) Data collection for hot place analysis, prediction

Data were extracted based on the population, business status, industry status, and economic category of the Seoul Data Plaza and public data portal. Table 1 shows the extracted data for each category. All data are based on district(Hang-Jeong-Dong).

Table 1. Extracted data by category

Population	Moving in, Moving out, single household, 2 households, 3 or more households, weekly population index, resident registration population
Industry Status	Food, Living Service, Leisure, Sports, Academic Education, Accommodation
Business	Operation Store / Closed Store
Economy	GDP, Level Index

III. Hot Place Selection and Data Visualization

Visualization of the number of Instagram tags and the number of Starbucks, McDonald's, and Mara soup stores by administrative district confirmed that several of the same regions stood out.

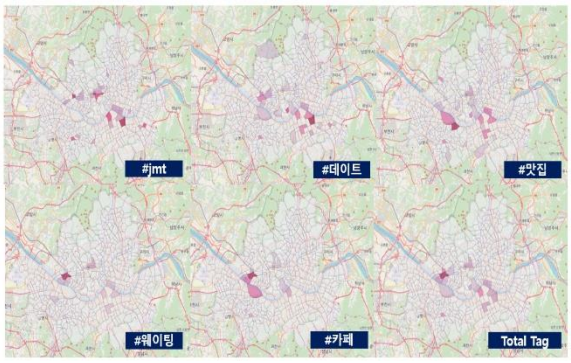

Figure 1. The number of each tag for each district

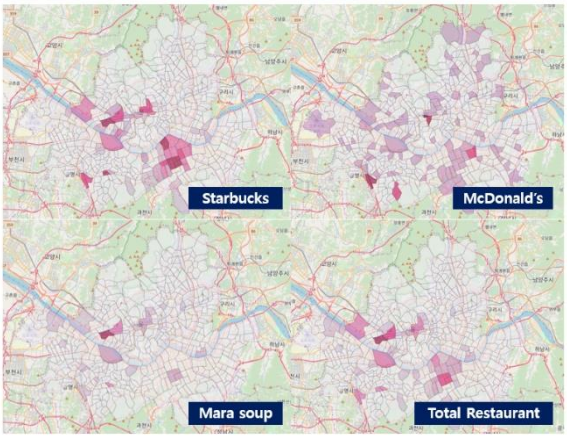

Figure 2. The number of each store (Starbucks / McDonald/Mara soup) for each district

A. *Hot Place Selection through Regional Scoring*

Total Mara soup stores are significantly smaller than Starbucks / McDonald's. Therefore, the number of Mara soup stores is multiplied by 2, and the number of Starbucks by region and McDonald's by region is added to score store density by region.

*Number of Marasoup stores per region * 2 + Number of Starbucks per region + McDonald's per region = Score of store by region*

The number of SNS posts per district was used to score regional SNS scores, which were then converted into total scores for each district using weights of 0.7 and 0.3(Because it's about selecting districts that 20 30 population "actually goes to")

SNS post count by district = SNS score by district
*SNS score by district * 0.7+ store score by district*0.3*
= Total score by district

As a result of examining the scores by the same process, we found that the score difference between the top 40 districts and the bottom 384 districts was large. Therefore, out of a total of 424 districts, 40 districts(Seogyo-dong, Sinsa-dong, Yeoksam 1-dong, Jongro 1,2,3,4 Ga …) are hot places. 384 districts(Burn 3dong, Suyu 3dong, ..)were classified as non-hot places.

IV. Hot Place Prediction Algorithm

A. *Variable Extraction*

Hot place being used as dependent variable, correlation coefficient for every data of each category was measured. As a result, the rest of the data were used as independent variables except for the transfer data with low correlation.

Table 2. Final variables

Dependent variable	Hot Place O/X
Independent variable	
Population	1-person households, 2-person households, 3-person households or more, Weekly Population Index, Registration Population
Industry Status	Food, Living Service, Leisure, Sports, Academic Education, Accommodation
Business	Operation Store/ Closed Store
Economy	GDP, Level Index

B. *Hot Place Prediction Algorithm*

Randomly classify administrative dataset into test and train set. Predict whether a test set is hot-placed using scikit-learn's multiple Logistic Regression. The average accuracy of the model was 0.91. At this time, the regions that has the actual value as 0 but has prediction result as 1 are extracted and assume that these areas are likely to become a future hot place. Repeat this process 1000times to count the extracted district, and select the top 10 districts

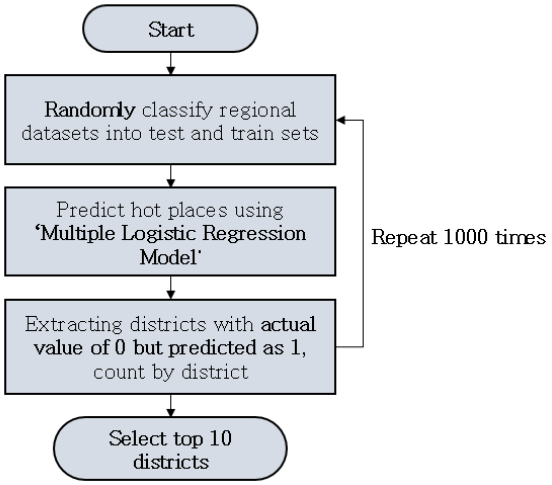

Figure3. Hot Place Prediction Algorithm

V. RESULTS

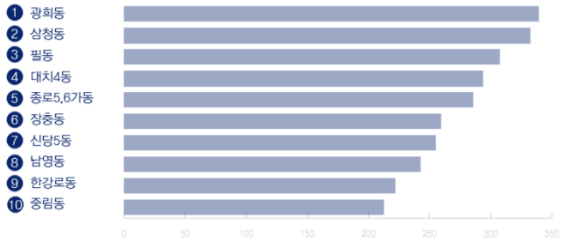

Figure 4. Top 10 districts

Out of the top 10 areas that are predicted by algorithm as a future hot place areas shown in the above figure, Junglim-dong, recently, was confirmed to have attracted many people through the opening of a multicultural space called Junglim Warehouse, with the recent influx of tourists due to the opening of elevated walkways. Furthermore, Namyoung-dong also attracts young business people from the printing plant area near Namyoung subway, where the trendy cafes and restaurants are located. In addition, it was confirmed that Pil-dong was selected as the alley road regeneration project in Seoul, and the alley road was starting to attract people due to the creation of Pil-dong Culture and Art Street. We can expect the following results through the hot place prediction algorithm shown in this paper.

VI. CONCLUSION

This paper predicted future hot places in Seoul. This would further be utilized for the following areas..

(1) Establishment of countermeasures against gentrification problems

In the cases of gentrification which are represented at neighborhoods of Hongdae, Seongsu-dong, and Gyeongnidan-gil, rent-fee and residents convenience problems have been constantly caused. As a result of this study, it is expected that the prediction of the ‘hot place’ can be made in advance to establish the policy and the countermeasures against it.

(2) Providing location insight to future entrepreneurs in their 20s and 30s

By directly or indirectly reflecting the culture that 20s and 30s like and lead, more specialized results could be obtained, thus providing location insights to future entrepreneurs preparing to start a business.

(3) Help local revitalization

By selecting highly correlated factors among the variables, we can provide insights to selected hot place candidates to help revitalize the region.

References

[1] J. E. Kim(2019), What is representative SNS by gender and age(2018), Retrieved Jan. 18. 2020, https://www.mk.co.kr/news/it/view/2018/03/174305.