[2019 Summer/Fall semester URP] KAIST URP Program

Individual Research Project Final Report

거시적 동선 정보에 기반한 고객 재방문 예측 및 장소 임베딩

Customer Revisit Prediction Using Macroscale Mobility Information and POI Embedding

산업및시스템공학과 김범영(20160101)

Contents

Abstract

Macroscale mobility information refers to expressing spatially and temporally the order in which customer moved. Wireless fingerprinting technology provides more accurate location information than GPS from inside to outside of the store. This enables to successfully predict revisit using only the data collected within specific store. At the same time, companies that have a broader range of macroscale mobility information also need for several analysis methods and predicting revisit. In this study, we propose the model for predicting customer revisit time. there are three main objectives. First, we present a model for predicting revisit time for specific categories or brand rather than a single store. Second, we solve the data leakage problem, which is not known whether the customer visit after that time when using data up to a certain point for training model. Third, we define features and embedding factors related to revisit in macroscale mobility information and to extend the WTTE-RNN methodology. As a result, Our model has less error in actual revisit time and can be classified well accordingly for customers who do not return than Cox, which is one of the existing models of survival analysis. Revisit prediction is highly applicable to various areas such as target marketing and customer group classification Moreover, we present pre-trained embedding to better characterize the data to develop the model in part 2. The data used in this study has a hierarchical structure So we use Poincare embedding for learning hierarchical representations by embedding entities into hyperbolic space. We make the assumption that The closer you are, the more similar categories or brands you have, the more frequently you visited at similar times, and the more similar complex you are in, the greater the similarity between POIs. Several examples show that the results of embedding follow assumptions. The embedding itself may be used as a method of quantifying the similarity between places. Furthermore, you can define user embedding based on the records that customers have visited in the past. This will eventually be able to classify customer groups and help target marketing.

Part 1 : Revisit Time Prediction Using Macroscale Mobility Information

1. Introduction

1.1 Research Background

Macroscale mobility information refers to expressing spatially and temporally the order in which customer moved. At this time, it is possible to identify the offline visiting behavior of the customer from semantic information such as what stores are located at that latitude and longitude. Wireless fingerprinting technology provides more accurate location information than GPS from inside to outside of the store, including the nearest store and the current floor. This enables to analyze funnels and hotspots in stores using offline customer movement records, and to successfully predict revisit using only the data collected within specific store.[1] At the same time, companies that have a broader range of macroscale mobility information also need for several analysis methods and predicting revisit. For example, Loplat X provides various services such as inducing payment by providing coupons or information to customer for nearby stores through providing location information services to client company's applications. In particular, the prediction of the next visit location and revisit using macroscale mobility information is highly applied to various areas such as target marketing and clustering customers. For example, if we can classify the customers to revisit a specific store, sales can be increased by providing an invitation to a new store nearby or providing a discount on a large purchase. Therefore, this study aims to predict the revisit time to a particular store through macroscale mobility information obtained through wireless fingerprinting technology.

1.2 Research Objectives

In this study, we propose the model for predicting customer revisit time. there are three main objectives.

Objectives 1) To present a model for predicting revisit time for specific categories or brand rather than a single store.

Targeting a particular single store is not only rare for customers who repeatedly visit it, but also for those places where a large number of customers have in common. This data scarcity is problem for learning and consequently, the service is not efficient. Therefore, we make a prediction based on the brand or category of the place. This not only solves the above problem to some extent, but also compares willingness to revisit with competitors in the same category

Objectives 2) To solve the data leakage problem, which is not known whether the customer visit after that time when using data up to a certain point for training model.

The data in this study represent customer movement over time. So if we use the cross validation, the test set could be earlier than the training set, which leads to data leakage. Therefore, the data should be divided according to time, and there is partial data that event does not occur within observation time. It makes to have large information loss and make biased prediction if we ignore that. So we use the survival analysis model.

Objectives 3) To define features and embedding factors related to revisit in macroscale mobility information and to extend the WTTE-RNN methodology.

We determine whether customer revisit to place category by using offline movement. So we have to define features and embedding factors representing each visit. And moreover, it is necessary to reflect the correlation with the past where the customer recently visited. For example, a customer who frequently goes to a cafe after going to a restaurant will be more likely to go to the cafe within a short time when he is currently in the restaurant. So we will use the deep learning RNN model that can reflect the order of data.

2. Related Research

2.1 Survival Analysis

Survival analysis is a branch of statistics that analyzes the time of occurrence of an event of interest, such as patient death or machine failure. The reasons for presenting the model using survival analysis are as follows. In applications with limited data collection periods, data scarcity issues exist, where the event does not occur within the observation period and the correct answer is not known. For example, in the case of using the data up to a specific point in time for the model training, the process of learning the model does not know whether the customer visits the point in time thereafter, and it is called censored data in the survival analysis. .A representative model of survival analysis is the Cox proportional hazards model[2]. The log-linear combination of covariates influences the occurrence of the event in a semiparametric way of determining when the event occurs. Furthermore, with the recent advances in machine learning and deep learning, there are Random Survival Forest[3] and Deepsurv[4] which express more complicated covariate relationship.

2.2 RNN

RNN is an abbreviation of Recurrent Neural Network. It is a kind of supervised learning among artificial neural networks in which hidden nodes are circulated. By forming a circular structure, the past data can be reflected in the sequence data, which is suitable for data processing such as voice and text. The figure below shows the basic (Vanilla) RNN structure, consisting of the input, the hidden state (layer), and the output. At this time, the state can be expressed as a circulating RNN cell because the structure of the learned parameters, W and Network, is applied equally in all steps.

$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$, $y_t = W_{hy}h_t + b_y$

The hidden state of the current state (t) is updated through the hidden state h_t and input of the previous point $h_{t=1}$ in time. The activation function used at this time is tanh.

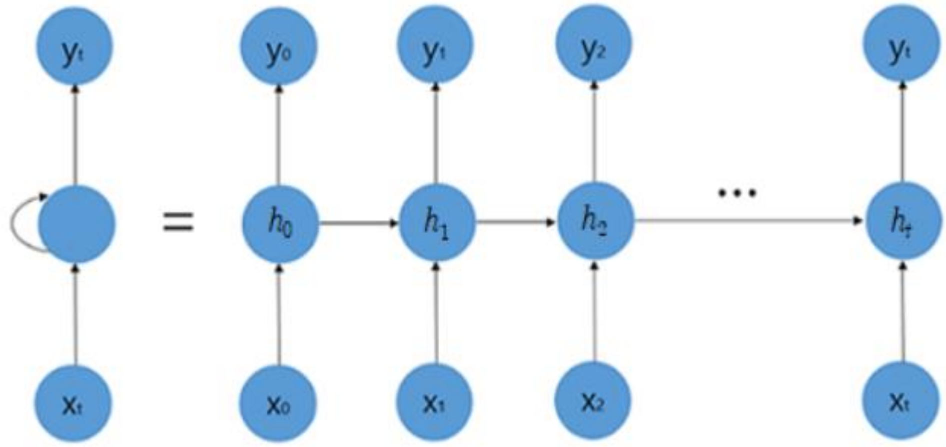

Figure 1 RNN Structure

Like other artificial neural networks, parameters (W, b) are updated in the direction of calculating forward propagation and reducing backpropagation errors through the label value of train data. In this case, the loss of the final network is defined as the sum of the loss of the output of each window step. But the problem with a basic RNN cell is the loss of distant information. If the distance from the past information is far away, the gradient gradually decreases in the course of back propagation, which reduces the learning ability. This is called long-term dependency. There are several models such as LSTM to solve this problem.

2.2 WTTE-RNN

The methods of survival analysis mentioned above can learn with censored data however, there are problems in that the characteristics related to the occurrence of the event must be proposed and the sequential data cannot be accurately reflected. To improve this, there are studies using the deep learning RNN algorithm can handle time varying covariates and learn temporal patterns and sequence of varying length.[5] In addition among them, the WTTE-RNN algorithm allows an intuitive analysis possible assuming a probability distribution function The WTTE-RNN uses the Weibull distribution,[6] which is used in probability function of event occurrence time such as machine failure. It is used everywhere for predicting things that will brake since It appears in nature like the normal distribution It can take many shapes by adjusting its two parameters α, β. The exponential and the discrete geometric distribution is the special case when β=1. This means that WTTE-RNN can include the exponential and Weibull Accelerated Failure Time model and the proportional hazards model that traditionally used in survival analysis. Therefore, this study extends the WTTE-RNN methodology by defining features and embedding factors related to revisit in macroscale mobility information to perform customer revisit prediction.

3. Methods

3.1 Problem Definition

As mentioned above, the macroscale mobility information refers to a set of location and time information of each place visited by each customer. When U is a customer, L is a store, and T is a time C as a category set, each visit v can be expressed as (u, l, t, c) ∈ U × L × T × C, which is the customer u. Indicates that he visited store l of category c at time t. At this time, several stores may be included in one category c, which may be expressed as follows. $\{l_1, l_2, \ldots l_k\} \subseteq$c. Given a particular place category c∈ C and time t∈

T, the goal is to find the time interval for the first return to the category after time t.

3.2 Data Description

To demonstrate the excellence of our model, we used data from 25,038 store visits collected by Loplat X over six months across Korea. This is a large enough logs for revisit predictions, and each store can be categorized into categories, with a total of 243, including coffee shops, restaurants, and libraries. Below is an example of a specific log.

Uid	Timestamp	Lat, Lng	Category	Brand
U1	2019-12-17 17:20:05	127.104765, 37.513579	Clothing Store	H&M

Specific Brand	Complex	Specific Complex	Address
H&M zamsil	Lotte World Mall	Lotte World Mall zamsil	300, Olympic-ro, Songpa-gu, Seoul

Table 1 Data description

3.3 Censored Data

When the observation period o^y for each customer is from the start to the end of the observation, customers may not visit certain categories within that period. This missing data is a problem for learning. At this time, we do not know exactly the time to visit a specific category c, but only after o^y. That is, the actual observation time y can be expressed like this. y^{u^*} is the actual event occurrence time for each customer. In the figure below, $y^u = o^u$ because it is censored data for blue, and $y^u = y^{u^*}$ because red is uncensored data.

$$y^u = \begin{Bmatrix} y^{u^*} & if \ \ uncensored \ \ data \\ o^u & if \quad \ \ censored \ \ data \end{Bmatrix}$$

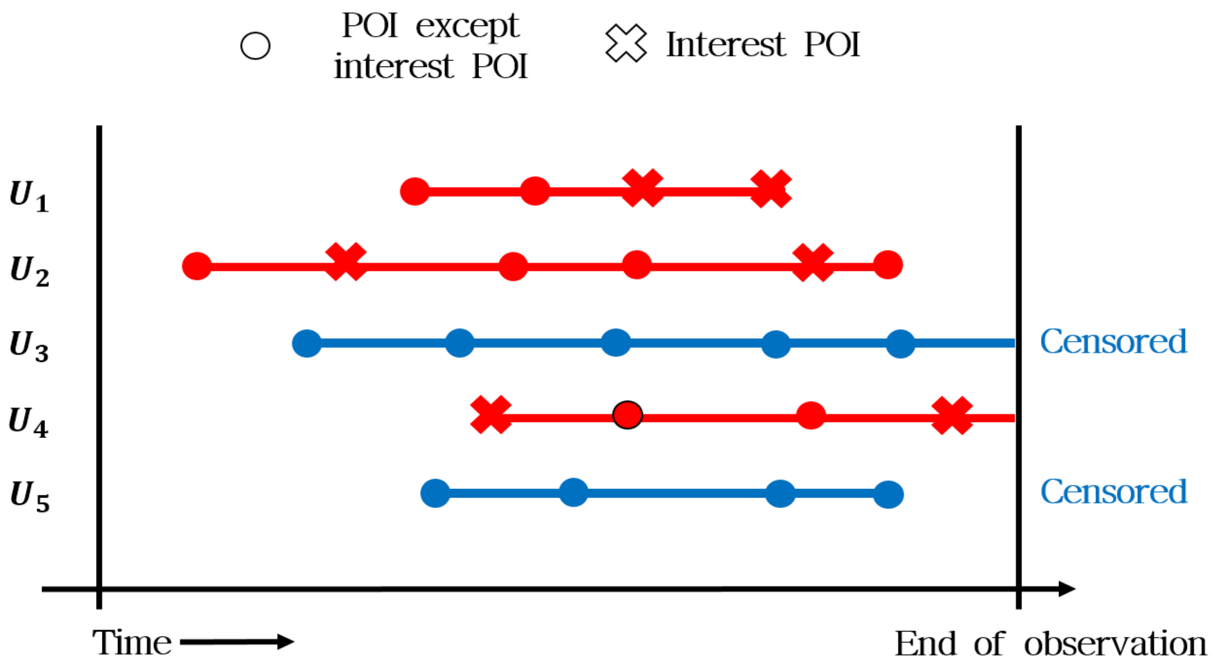

Figure 2 Macroscale mobility information and censored data

3.4 Loss Functions

Loss function is set differently depending on whether censored or not. It was designed with reference to WTTE-RNN. For uncensored data with a definite label visited in the place category, the probability of occurring at that time is maximized. In the case of censored data that does not return within the observation time, we learn to maximize the probability that an event will occur after the last visit except the target place category of interest. y_t^n is the event time for user n = 1,⋯ , N at time step t = 0, 1,... T_n $x_{0:t}^n$ is data up to time t and u_t^n is indicating censored

$$\sum_{n=1}^{N}\sum_{t=0}^{T_n} u_t^n \log[\Pr(Y_t^n = y_t^n | x_{0:t}^n) + (1-u_t^n)\log[\Pr(Y_t^n > y_t^n | x_{0:t}^n)]$$

$$u = \begin{Bmatrix} 1 \;\; if \;\; uncensored \;\; data \\ 0 \;\; if \;\;\;\;\;\; censored \;\; data \end{Bmatrix}$$

3.5 Model

The assumption is that the revisit time probability distribution function follows the Weibull distribution. The Weibull distribution consists of two variables α and β. When expressed in the discrete form of the Weibull distribution, the loss function described above can be expressed like this.

$$\mathbf{L} = \mathbf{log}(\boldsymbol{p}(\boldsymbol{y})^{\boldsymbol{u}}\boldsymbol{S}(\boldsymbol{y}+\mathbf{1})^{\mathbf{1}-\boldsymbol{u}}) = \boldsymbol{ulog}\left(\boldsymbol{e}^{\left(\frac{y+1}{\alpha}\right)^{\beta}-\left(\frac{y}{\alpha}\right)^{\beta}}-\mathbf{1}\right)-\left(\frac{\boldsymbol{y}}{\boldsymbol{\alpha}}\right)^{\boldsymbol{\beta}}$$

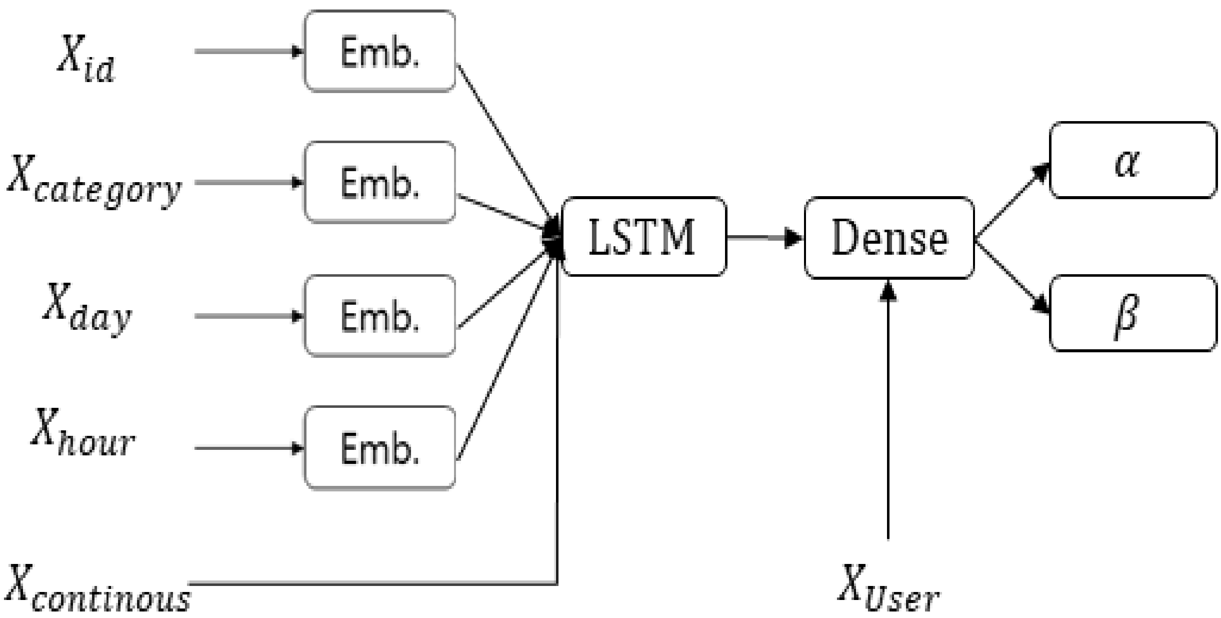

Figure 3 The simplified architecture of Model

In this study, we used LSTM architecture designed to solve the long-term dependency problem on the basic RNN structure. At this time, in order to learn the variables α and β, both the features generated at each visit and the features of the entire period are used. In the figure, the embedding features ($X_{category}$, X_{day}, X_{hour}, X_{id}) and the continuous feature ($X_{continuos}$) are the features that represent each visit to learn the sequence of movement through the RNN. At this point, the visits from the last to a certain point of time are only selected and embedded assuming that the impact is greater in the near past. The number of past visits to embed can be adjusted experimentally. X_{user} represents the overall characteristics of the entire period. The features used are as follows.

- Embedding Features ($X_{category}, X_{day}, X_{hour}, X_{id}$)

1) Category 2) Day of the week 3) Hour 4) Customer Id

- Continuous Features ($X_{continuos}$)

1) Time until next any log

- Aggregated Features (X_{user})

* Also used in Cox Model

1) The number of interest category over the entire period

2) The number of all logs

3) The number of days with logs

4) The number of all logs divided by the days with logs

5) Time from last visit of interest category

6) The number of the weekend over the entire period

7) Average interval between interest category

8) Average interest between all logs

4. Results and Conclusion

4.1 Measures

The performance of predicted revisit time is compared with the Cox algorithm that can be used for learning censored data. Root Mean Square Error (RMS), C-index[7] (Concordance Index), Non-returning Recall, and Non-returning F1 score are used as the measurement. First, root mean square error is the most commonly used evaluation criterion in general regression problems and represents the degree of error between the predicted and actual values. This can only be calculated if there is revisit. In fact, it is the magnitude of the error of the uncensored data. Secondly, C-index is the most commonly used evaluation standard in survival analysis. It looks like this.

$$C = \frac{1}{|\varepsilon|} \sum_{T_i\ uncensored} \sum_{T_j > T_i} 1_{f(x_i) < f(x_j)}$$

$|\varepsilon|$ is the number of instances and f(x) is the predicted event occurrence time. It is the degree of similarity between the ranking of the actual event time and the ranking of the predicted value. Finally, the non-returning recall and non-returning F1 scores are the criterion for evaluating the accuracy of binary classification, given that there is no actual visit value within a given observation period. The revisit model should not only predict well similar to the actual value if there is a revisit within the observation period, but if not, it should be predicted beyond the observation period to determine that the revisit is not within that period.

4.2 Results of "Coffee Shop"

In this study, the learning was conducted using the total period as one. (T_n=1)

	Cox	WTTE-RNN
C-index	0.698	**0.726**
RMSE	205.24	**199.49**
non-returning recall	**0.488**	0.438
non-returning f1	0.543	**0.555**

Table 2 Results of 'Coffee Shop'

This is the result of experiment based on the category of 'Coffee shop'. Our model has less error in actual revisit time and can be classified well accordingly for customers who do not return than Cox, which is one of the existing models of survival analysis.

4.3 Conclusion

Revisit prediction is highly applicable to various areas such as target marketing and customer group classification, so there is a need for research. In this study, we solved the data scarcity problem through survival analysis and reflected the order of visits through the RNN model. In addition, we extended the WTTE-RNN methodology by defining features and embedding factors related to revisit in macroscale mobility information, and successfully predicted revisit of customers.

5. Future Research Plan & Proposal

5.1 Future Research Plan & Proposal

In the case of data provided by Loplat X, the location information of the customer is automatically recorded as a log at a certain time interval, and the place where the user went, not the place where the customer actually visited, will be recorded. Therefore, by preprocessing these data, we can create a model that can be predicted quickly without sacrificing performance using only meaningful visits. In addition, refinement of the features that feed the model is required. If we can obtain customer and place embedding that reflect spatial and temporal information well from macroscale mobility, good results can be obtained from the order of visits without comprehensive features such as X_{user}.

Part 2 : POI Embedding

1. Introduction

1.1 Research Background

In Part 1, we show successful results for predicting customer revisit. We will then present pre-trained embedding to better characterize the data to develop the model. Embedding is an necessary process for computers to understand natural language in text-based models. Embedding can express semantic similarities between words and is more efficient than one-hot vectors in terms of memory. In this study, there are texts that need to identify relationships such as categories and brands. Also, the embedding itself may be used as a method of quantifying the similarity between places and further measuring the similarity between users according to which place they visited.

1.2 Research Objectives

The data used in this study has a hierarchical structure. For example, Starbucks Jamsil is included in Starbucks and Starbucks is included in a category coffee shop. There is a need for an embedding methodology that understands this hierarchical relationship. So we made the following assumption to make embedding related to structure of data we use.

Assumption) The closer you are, the more similar categories or brands you have, the more frequently you visited at similar times, and the more similar complex you are in , the greater the similarity between POIs.

2. Related Research

2.1 Poincare Embedding[8]

Representation learning has become an important approach in machine learning and artificial intelligence. For instance, word embedding such as word2vec, Glove are widely used for tasks ranging from machine translation to sentiment analysis. In the Poincare embedding, they study the influence of the underlying geometry on embedding structured data. However, Representing a hierarchical structure in Euclidean space requires large dimensions which, in turn, causes significant problems with regard to the computational complexity. So they work on hyperbolic embedding that is novel approach for learning hierarchical representations by embedding entities into hyperbolic space.

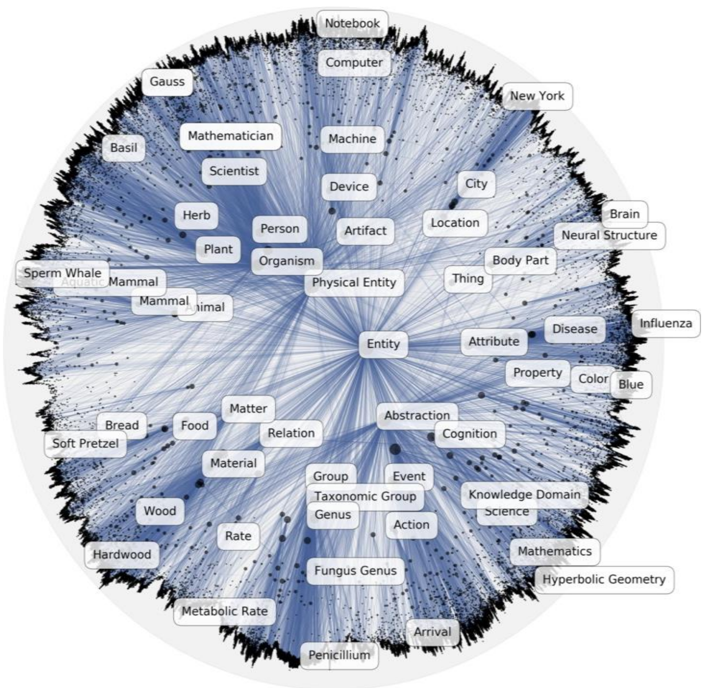

Figure 4 Poincare embedding

3. Methods

3.1 Main Features of Hierarchical

In order to learn Poincare embedding, we need to show the tree relation of poi data. In this study, the following features were used. The visualization of the structure is below

1) hierarchical category and brand
2) geometrical similarity
    - Address cluster
    - Latitude and Longitude cluster
3) effect of time zone
4) effect of complex

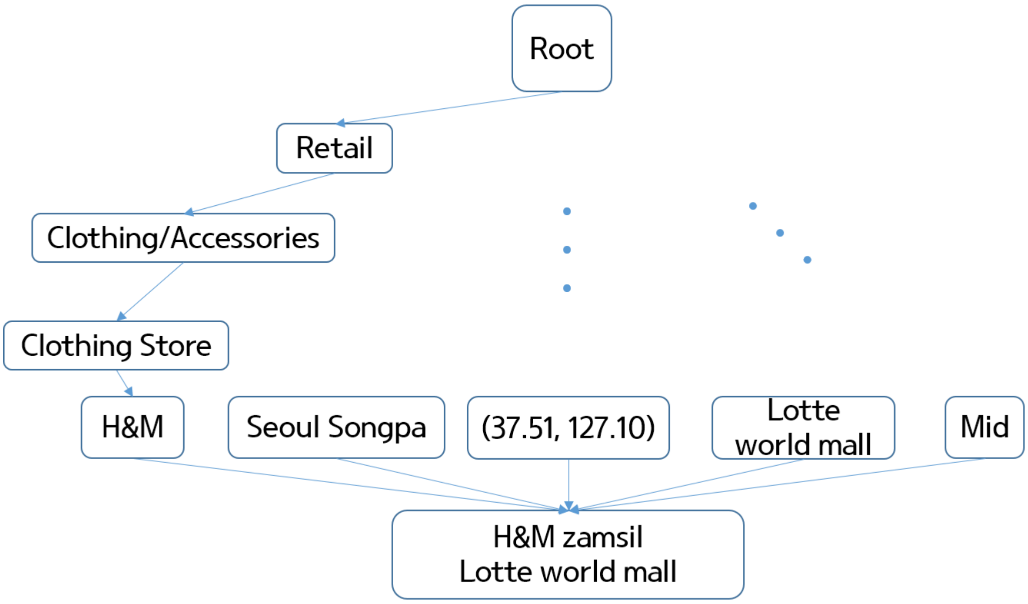

Figure 5 Structure of data

4. Results & Conclusion

4.1 Results

Similarity could be judged by Poincare distance and cosine similarity. In these examples, I use the cosine similarity.

Example 1) Category sets with the same parent are more similar

Cosine similarity	Snack Bar	Costmetics/Perfumes
Restaurants	0.0649	-0.0165

Cosine similarity	Middle School	Pharmacy/Medical Equipments
College Education	0.886	0.315

Table 3 Cosine similarity of first example

Example 2) When located in the same complex, poi having similar categories are considered similar

Cosine similarity	Mad for Garlic Lotte cinema world tower mall	H&M Men Lotte wolrd mall
H&M zamsil Lotte world mall	0.0162	0.2811

Cosine similarity	Seorae Cold Noodles Lotte world mall	Unmanned Goods Lotte world mall
Mad for Garlic Lotte cinema world tower mall	0.5107	0.1915

Table 4 Cosine similarity of second example

Example 3) When a place is fixed, similar brands or categories are similar.

Top Cosine similarity between Paris baguette Mapo Yeom-ri and poi in 'Seoul Seodaemun-gu'

Paris Baguette Mapo Yeomli-Shop	Cosine similarity
Paris Baguette Chongjeong Diovil-Shop	0.6645
Paris Baguette Seodaemun Happy	0.5880
Panpane Seodaemun	0.5736
Nubake Seodaemun	0.5622
…	…

Example 4) Close poi is similar when a brand or category is fixed

Cosine similarity between Lotteria and McDonald 's Daejeon DT

McDonald's Daejeon DT	Cosine similarity
Lotteria Daejeon eunhaeng	0.6792
Lotteria Daejeon Hannamdae	0.1971
…	…
Lotteria Seosan Lake Park Branch	-0.3188

4.2 PCA & T-SNE

PCA and T-SNE methods were used to visualize high dimensional embedding vectors. As a result, we could see places that are located nearby or have similar categories or brands are more similar. Below is the T-SNE of category embedding.

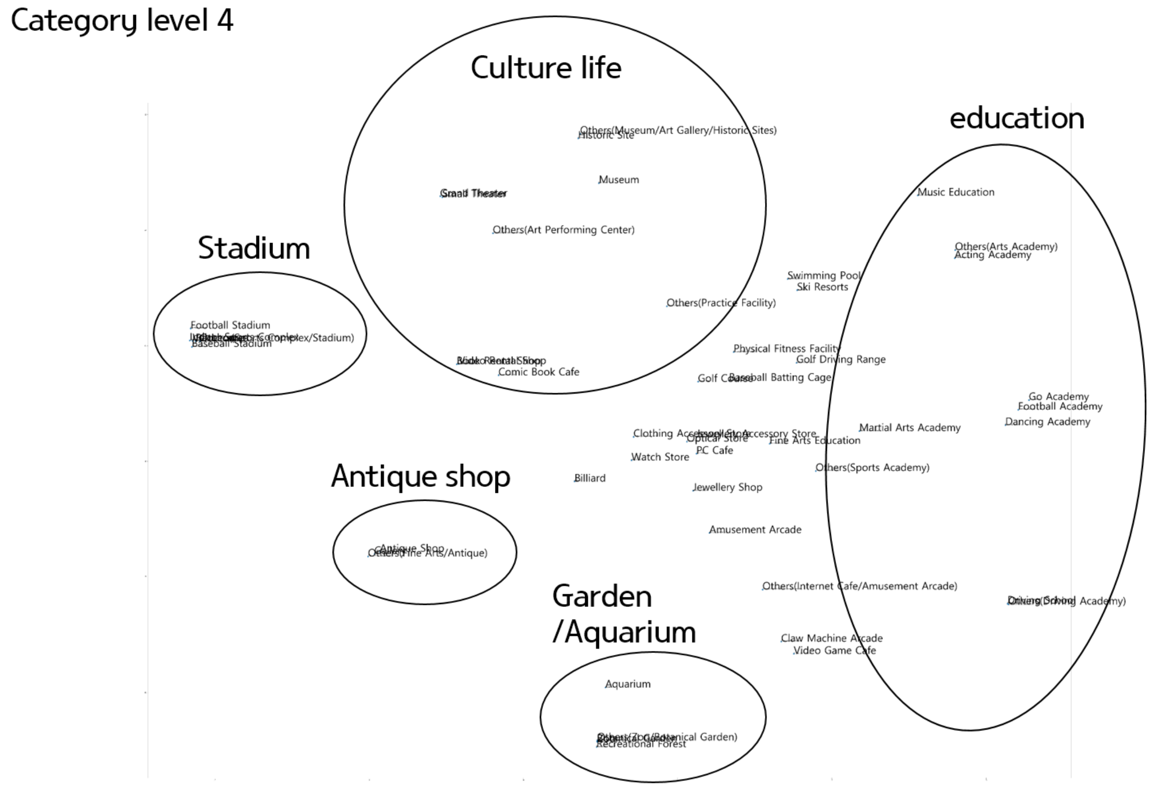

Figure 6 T-SNE of category

5. Future Research Plan & Proposal

5.1 Future Research Plan & Proposal

Refinement of POI embedding is needed. For data that we have, there are a lot of missing values and we don't know what location it stands for. Data preprocessing will be necessary using external data such as geo mapping. Also it is needed to reinforce the features that represent the place. If you construct a complex tree by further subdividing location or time zone information, you will get more realistic results. And furthermore, you need to define user embedding based on the records that customers have visited in the past. This will eventually be able to classify customer groups and help target marketing.

6. Reference Literature

[1] Sundong Kim and Jae-Gil Lee,"Utilizing In-store Sensors for Revisit Prediction,"*IEEE International Conference on Data Mining*, pp. 217-226, 2018.

[2] David Roxbee Cox,"Regression Models and Life Tables (with Discussion),"*Journal of the Royal Statistical Society: Series B*, Vol. 34, No. 2, pp. 187-220, 1972.

[3] Hemant Ishwaran et al.,"Random Survival Forests," *The Annals of Applied Statistics,* Vol. 2, No. 3, pp. 841-860, 2008.

[4] Jared Lee Katzman et al., "DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network,"*BMC Medical Research Methodology,* Vol. 18, No. 1, p. 24, 2018.

[5] Guolei Yang et al.,"Spatio-Temporal Check-in Time Prediction with Recurrent Neural Network based Survival Analysis,"*International Joint Conference on Artificial Intelligence*, pp. 2976-2983, 2018.

[6] Egil Martinsson, "WTTE-RNN : Weibull Time to Event Recurrent Neural Network,"Chalmers University of Technology, Gothenburg, Sweden, 2016.

[7] Vikas Chandrakant Raykar et al., "On Ranking in Survival Analysis: Bounds on the Concordance Index,"*Advances in Neural Information Processing Systems 20*, pp. 1209-1216, 2008.

[8] Maximilian Nickel et al., “Poincare Embeddings for Learning Hierarchical Representations,” *Advances in Neural Information Processing Systems 30*, pp. 6339-6347, 2017.