Capstone Project

A location content-aware recommender system

Introduction

Newly emerging location-based and event-based social network services, such as Foursquare provide us with a new platform to understand users preferences based on their activity history, and thus, we are able to recommend event or/and venues based on the users preferences.

For this capstone data science project i decided to implement a recommender system. There many different algorithm to do this,on of the most used are traditional collaborative filtering-based recommender systems. But, this method is limited since a user can only visit a limited number of venues/events and most of them are within a limited distance range, so the user-item matrix is very sparse.The problem becomes more challenging when people travel to a new city where they have no activity history.

LCARS can alleviate these problems by considering both user preferences and local preferences into recommendation.

Summary

1. Presentation of LCARS algorithm
2. Presentation of the data set
3. Data processing
4. Results
5. References

1. LCARS

1.1. Presention of LCARS

LCARS, a Location content-aware algorithm recommender system that offers a particular user a set of venues (e.g., restaurants) or events (e.g., concerts and exhibitions) by giving consideration to both personal interest and local preference. This recommender system can facilitate people’s travel not only near the area in which they live, but also in a city that is new to them.

1. LCARS

1.1. The architecture Framework of LCARS

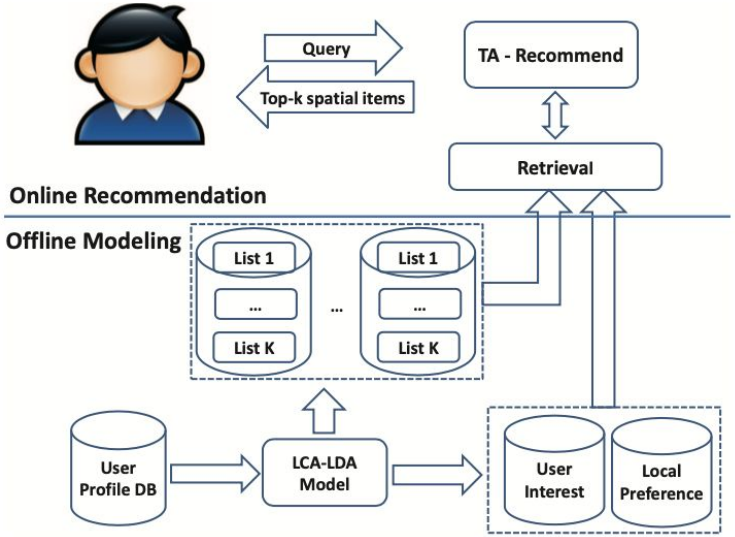


Figure 1: The Architecture Framework of LCARS

1. LCARS

1.1. The architecture Framework of LCARS

LCARS consists of two components: offline modeling and online recommendation. The offline modeling part, called LCA-LDA, is designed to learn the interest of each individual user and the local preference of each individual city by capturing item co-occurrence patterns and exploiting item contents. The online recommendation part automatically combines the learnst interest of the querying user and the local preference of the querying city top produce the top-k recommendations.

1. LCARS

1.2. Model description

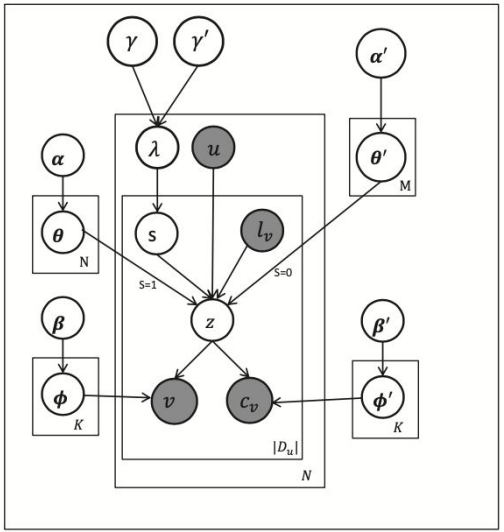


Figure 2: The Graphical Representation of LCA-LDA

SYMBOL	DESCRIPTION
N, V, M, C	the number of users, spatial items, locations, content words
$\mathcal{U}, \mathcal{V}, \mathcal{L}, \mathcal{C}$	the set of users, spatial items, locations, content words
$\mathcal{V}_l$	the set of spatial items located in location l
K	the number of topics
D_u	the profile of user u
v_{ui}	the spatial item of ith record in user profile D_u
θ_u	the interest of user u, expressed by a multinomial distribution over topics
θ'_l	the local preference of location l, expressed by a multinomial distribution over topics
ϕ_z	a multinomial distribution over spatial items specific to topic z
ϕ'_z	a multinomial distribution over content words specific to topic z
z_{ui}	the topic assigned to spatial item v_{ui}
l_{ui}	the location of spatial item v_{ui}
l_u	the querying location of the querying user u
c_{ui}	a content word describing spatial item v_{ui}
$\mathcal{C}_{ui}$	the set of content words describing spatial item v_{ui}
s_{ui}	if spatial item v_{ui} is generated by θ_u or $\theta'_{l_{ui}}$
β, β'	Dirichlet priors to multinomial distributions ϕ_z, ϕ'_z
α, α'	Dirichlet priors to multinomial distributions θ_u, θ'_l
λ_u	the mixing weight specific to user u; the parameter for sampling the binary variable s
γ, γ'	Beta priors to generate λ_u

Table 1: Notations used in the paper

1. LCARS

1.2. Model description

where $P(v|\theta u,\varphi,\varphi)$ is the probability that spatial item v is generated according to the personal interest of user u, denoted as θu, and $P(v|\theta', \varphi, \varphi')$ denotes the probability that spatial item v is generated according to the local preference of lu, denoted as θlu . The parameter λu is the mixing weight which controls the motivation choice

2. Data set presentation

The data set on which the model was learned consists of 161,468 rows, each row represent a user interest (= a checkin) of 646 users , from three big cities : Chicago, New York City and Boston .

	user_id	venue_id	category	city
0	180962	4b3be5b9f964a520e37d25e3	Bar	Boston
1	38722	4b3be5b9f964a520e37d25e3	Bar	Boston
2	93711	4b3be5b9f964a520e37d25e3	Bar	Boston
3	68294	4b3be5b9f964a520e37d25e3	Bar	Boston
4	101835	4b3be5b9f964a520e37d25e3	Bar	Boston

Some numbers :
Average number of users per user : 13.6 (Chicago), 12.1 (Boston), 10.2 (New York)
Number of users :
Number of venues :

This dataset was constructed from a larger data set that includes long-term (about 18 months from April 2012 to September 2013) global-scale check-in data collected from Foursquare and contains 33,278,683 checkins by 266,909 users on 3,680,126 venues (in 415 cities in 77 countries). Those 415 cities are the most checked 415 cities in the world, each of which contains at least 10000 check-ins),

2. Data set presentation

It contains three files in tsv format.

- File dataset_TIST2015_Checkins.txt contains all check-ins with 4 columns, which are:
1. User ID (anonymized)
2. Venue ID (Foursquare)
3. UTC time
4. Timezone offset in minutes (The offset in minutes between when this check-sin occurred and the same time in UTC, i.e., UTC time + offset is the local time)

- File dataset_TIST2015_POIs.txt contains all venue data with 7 columns, which are:
1. Venue ID (Foursquare)
2. Latitude
3. Longitude
4. Venue category name (Foursquare)
5. Country code (ISO 3166-1 alpha-2 two-letter country codes)

- File dataset_TIST2015_Cities.txt contains all 415 cities data with 6 columns, which are:
Venue category ID (Foursquare)
1. City name
2. Latitude (of City center)
3. Longitude (of City center)
4. Country code (ISO 3166-1 alpha-2 two-letter country codes)
5. Country name
6. City type (e.g., national capital, provincial capital)

2. Data set processing

2.1. Pre-processing The code for pre-processing is available in the notebook :LCARS-data-pre-processing

Data pre-processing includes the steps from the extraction of the data from the original data set to the the creation of a data set on which we are going to learn LCA-LDA model. The data pre-processing consists in the followings steps :

1. Start with import venues data set (dataset_TIST2015_POIs.txt), and keep the venues from US only.
2. Import cities data set and (dataset_TIST2015_Cities.txt), and use the algorithm of the k nearest neighbors with *scikit learn* to assign to each venues the nearest city from its latitude and longitude.
3. Arbitrary select 3 cities, in our case they are Boston, Chicago and New York cities, and for each of these cities we create a *pandas* dataframe, and merge with the checkins dataset (dataset_TIST2015_Checkins.txt) for each one. Thus we have our user profile dataset. But it is not finished yet.
4. for simplicity's sake, i decided, for user that have several checkins of just one venue, to keep only one occurence of each venue.
5. The data set is still too large, so i reduced it by selecting a limited and equal number (4500) of users in each city.
6. Our user profiles data set is now builded, so now we have to do some wrangling processes on it before being ready for the model

2. Data set processing

2.2. Data Wrangling

Notebook : LCARS.ipynb (PART 1 : Initialization)

1. From user profiles data frame, we construct user profiles matrix, with tokenized and stemmed content words
   Library : NLTK

2. Encode user profiles matrix by replacing each value with a simple Id.

3. Experimental Result

The evaluation method is described in the paper section 3.1.3.
I selected 3 user profiles :

--- User Profile (uid : 1797) ---
User from Chicago travel to New York :query(1797,nyc)
Number of checkins in up2 : 148
Nb of checkins in NYC : 4
Nb of checkins in Chicago : 144
Nb of checkins in Boston : 0

-- User Profile (uid : 1583) ---
User from NYC travel to Chicago : query(1583,chicago)
Number of checkins in up2 : 101
Nb of checkins in NYC : 89
Nb of checkins in Chicago : 12
Nb of checkins in Boston : 0

--- User Profile (uid : 2954) ---
User from Chicago travel to NYC : query(2954,nyc)
Number of checkins in up2 : 125
Nb of checkins in NYC : 8
Nb of checkins in Chicago : 117
Nb of checkins in Boston : 0

For each of these user profiles, we divide a user's activity history into a test set and a training set under two different dividing strategies :

1 - For the first strategy, we select all spatial items visited by the user in a non-home city as the test set and use the rest of the user's activity history in other cities as the training set.

2 - For the second setting, we randomly select 20% of spatial items visited by the user in personal home city as the test set, and use the rest of personal activity history as the training set.

According to the above designed dividing strategies, we split the user activity history S into the training data set Straining and the test set Stest. To evaluate the recommender models, we adopt the testing methodology and the measurement Recall@k

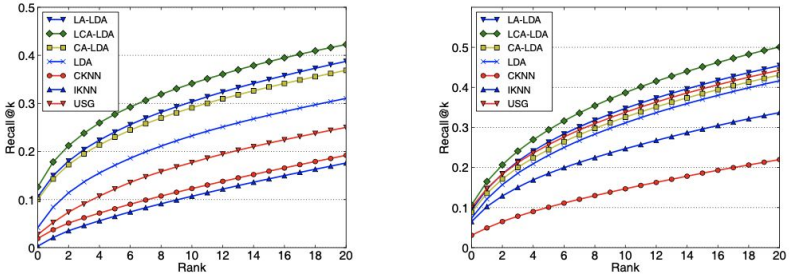
3. Experimental Result

number of topic : 50
k (top k result) = 20

User Profile (uid)	Recall (querying cities are new to querying users)	Recall (querying cities are the home cities of querying users)
1797	0.5	0.25
1583	0.25	0.16
2954	0.63	0.20
Average=	0.46	0.20

3. Experimental Result

We compare the results of our experiments with those of the original paper.
As shown in Figure 3(a) where querying cities are new cities, the recall of LCA-LDA is about 0.42 when k = 20 (i.e., the model has a probability of 42% of placing an appealing event within the querying city in the top-20).
It is a better result than ours. Because, on the DoubanEvent dataset is larger and includes content informations. foursquare dataset includes only cagegory name.
Taking these informations into consideration, we can conclude that the implementation of LCARS over a sample from foursquare dataset is a success.



(a) Users Traveling in New Cities (b) Users Traveling in Home Cities

Figure 3: Top-k Performance on DoubanEvent

4. References

LCARS: A Location-Content-Aware Recommender System, Hongzhi Yin Yizhou Sun, Bin Cui† Zhiting Hu, Ling Chen. https://www.cs.cmu.edu/~zhitingh/data/kdd13_lcars.pdf

A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling William M. Darling School of Computer Science University of Guelph. http://u.cs.biu.ac.il/~89-680/darling-lda.pdf

The data set was formed by, Yang, Dingqi and Zhang, Daqing and Zheng, Vincent. W. and Yu, Zhiyong, for the article *Modeling User Activity Preference by Leveraging User Spatial Temporal,* in the journal IEEE Transactions on Systems, Man, and Cybernetics: Systems},

Notebooks :
data pre processing :
https://github.com/cocoftn/capstone-project/blob/master/LCARS-Data-Collecting.ipynb
main : https://github.com/cocoftn/capstone-project/blob/master/LCARS.ipynb