Context-Aware Implicit Feedback based Hotel Recommender System for Anonymous Business Travellers

Molood Arman

Supervisor: Nacera Bennacer

Advisors: Srudeep Katamreddy
Christophe Blaya

Confidential Thesis

2018

Agenda.

01. Introduction

02. Literature Review

03. Preprocessing & Analysis Data

04. Solution

05. Evaluation

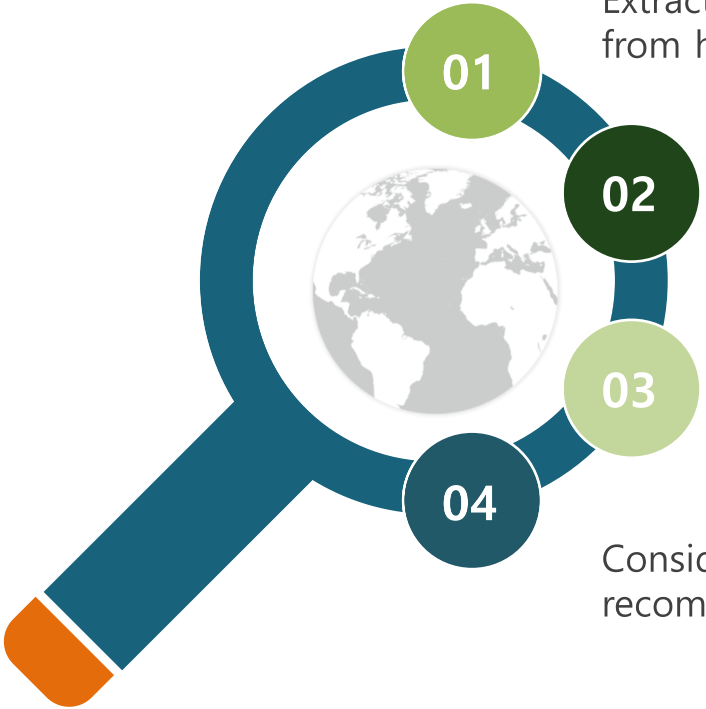
06. Conclusion & Future works

06. Q & A

Problem Statement

Booking data is gathering
From different travel agencies

Objectives



01 Extract business policy of each Company from historical data

02 Developing a model based recommender system (RS) based on historical data

03 Improving RS performance by embedding profiles created just based on historical data

04 Considering context to give more accurate recommendations to each customer.

Evolution Of RS

Item Hierarchy

If you buy printer, you will need ink

-BestBuy

Attribute-based RS

You like action movies starring Clint Eastwood, then you probably like "good, bad and the ugly"

-Netflix

Collaborative Filtering (user-user similarity)

People like you, who bought beer, also bought diapers.

-Target

Collaborative Filtering (item-item similarity)

You like Godfather, then you will like Scarface

-Netflix

Model based

training SVM, LDA, SVD, NN and etc. for implicit features

Top N-Recommendations
Memory based

Challenges

Sparsity

There is no access to enough information

Cold Start

There are not any current available data for new items /users

Scalability

Capacity to deal with growing amount of data.

Over Specialization

Users are limited to getting recommendations which definitely known or defined in their profiles

	Companies	Hotels	Density
Data	38,308	94,074	0.025%

Context Aware RS

Contextual Pre Filtering

2D recommender system

The current context is used for just selecting only the relevant set of data

Contextual Post Filtering

2D recommender system

The resulting set of recommendations is adjusted for each user using the contextual information

Contextual Modeling

MD recommender system

contextual information is used directly inside the model as part of the rating estimation

Data Acquisition

01 **Explicit Data**
customer rating or feedbacks

02 **Implicit Data**
Purchase History

03 **Side Information**
product Attributes or Product Descriptions

Data Acquisition

```
> er

--- RLR ---
RP/NCE1A016U/NCE1A016U            FJ/SU  14FEB17/1735Z   3TOC3B
  1.FANI/JEAN MR
  2  AF6203 Y 20JUN 2 NCEORY HK1  0630 0755  20JUN  E  AF/3TOC3B
  3 AP 999999999999
  4 TK TL15FEB/NCE1A016U
  5 OPW-24FEB:1100/1C7/AF REQUIRES TICKET ON OR BEFORE
        27FEB:1100/S2
  6 OPC-27FEB:1100/1C8/AF CANCELLATION DUE TO NO TICKET/S2
  7 RM TEST AMADEUS
  8 RM TEST
```

Passenger Name Record (PNR)

Preprocessing & Analysis Data

Raw **Data** → Name **Companies** → Preprocessing **Data** →

Duplication → **Transforming Data** → **Missing Value**

Introduction
Literature Review
Data preprocessing
Solution
Evaluation
Conclusion

	Type	Distinct Value	Missing Value
Date	String	2405659	0
PropertyId	String	94314	0
ChainCode	String	319	0
RateCode	String	11455	0
RoomType	String	4679	0
LeadTime	String	458	0
officeId	String	5911	0
rooms	Integer	2	67
nights	Integer	98	0
guests	Integer	7	52356
adults	Integer	7	0
SupplierCode	String	3	2941578
TotalPriceEuro	Integer	259669	1730124
BaseamountEuro	Integer	333424	562074
CommisionEuro	Integer	227769	1839625
MarkUp	Integer	1781	2942875
email	String	1044875	0

Date	String	2405659	0
PropertyId	String	94074	0
ChainCode	String	319	0
RateCode	String	11455	0
RoomType	String	4676	0
LeadTime	Integer	361	0
officeId	String	5911	0
Hotelname	String	94074	0
nights	Integer	98	0
Citycode	String	4106	0
CityName	String	17018	0
SupplierCode	String	3	0
PricePerNR	Integer	27	0
TRavelNet	String	1356	0
CountryCode	String	202	0
Month	String	7	0
Comapny	String	38308	0
Commisioneuro	Integer	23	0

Inferring Business Policy

Representative Context

Frequent itemset Mining

Needs lots of memory

Association Rule Mining

Apriori Algorithm

Support

Confidence

Lift

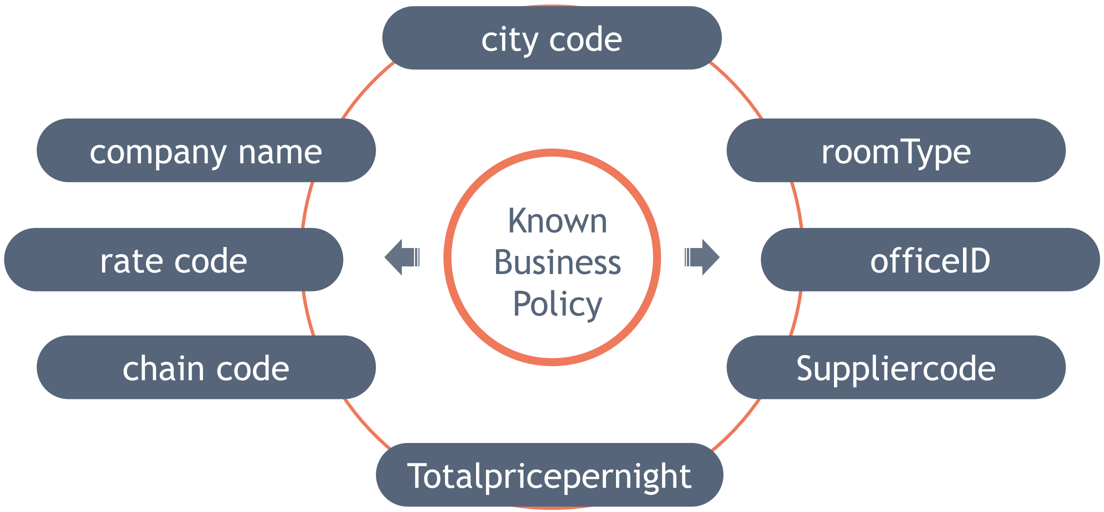
Pattern	Frequency
MCKINSEY|MUCAP21WH|MCK|ZQ	9002
NoCompany[illegible]	[illegible]
EDF[illegible]	[illegible]
BABCOCKINTE[illegible]	[illegible]
NORDEA[illegible]	[illegible]

Diversity of Itemset
31485
21946
[illegible]
[illegible]
28193

items=frozenset('DEF', 'ROH', 'MILAX21BN', 'HS', 'FAT', 'FCAGROUP','5'), support=0.07995707002951435, confidence=1.0, lift=4.781270044900578

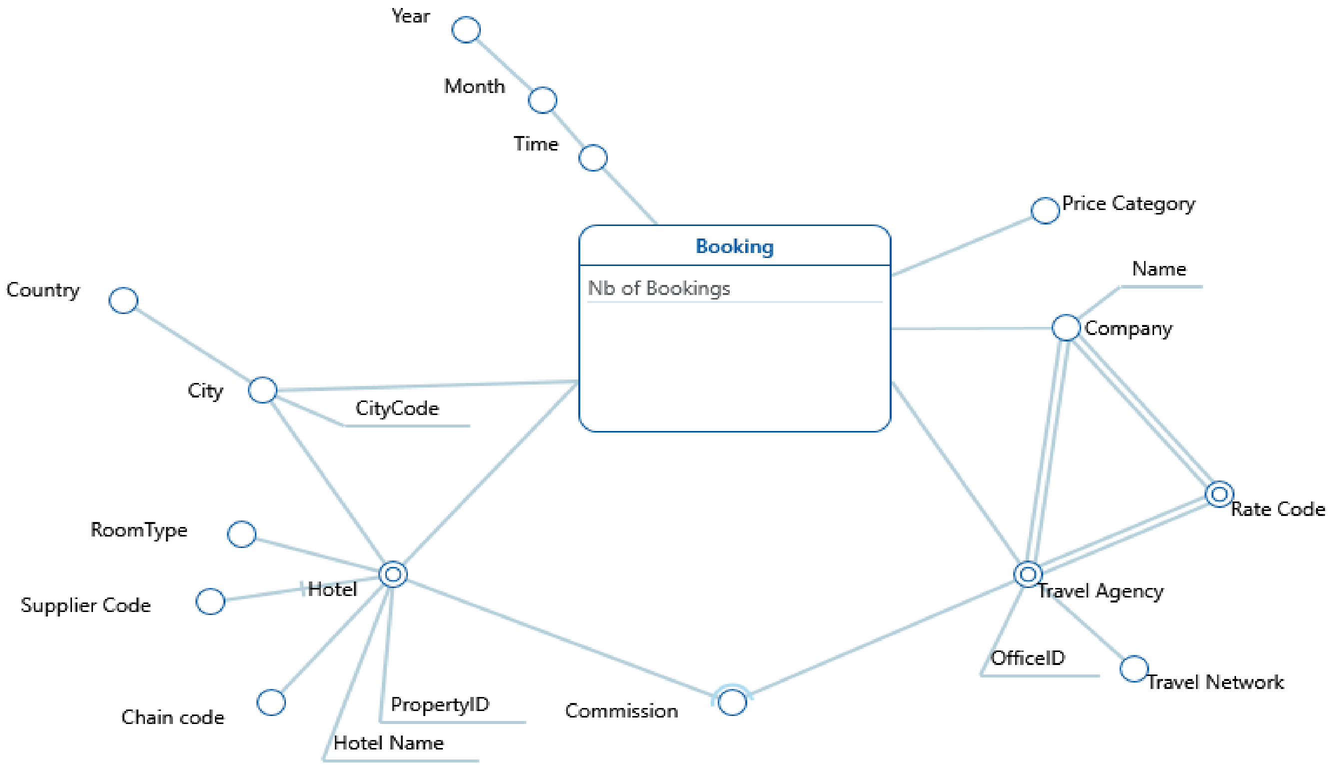
Inferring Business Policy

Representative Context



Integration OLAP with RS

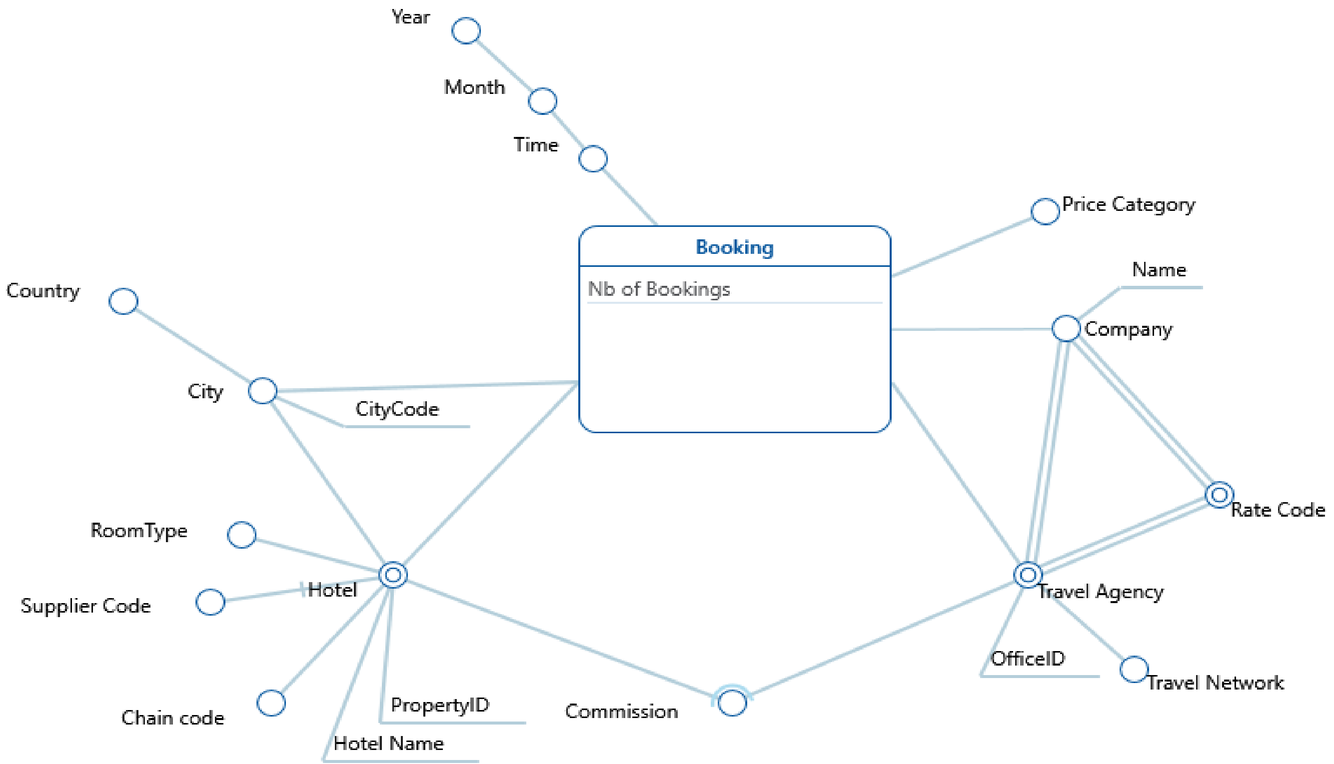
Multi Dimensional Model



Introduction
Literature Review
Data preprocessing
Solution
Popularity base RS
Similarity Base RS
Matrix Factorization
Evaluation
Conclusion

Integration OLAP with RS

Multi Dimensional Model



Integration OLAP with RS

Multi Dimensional Model

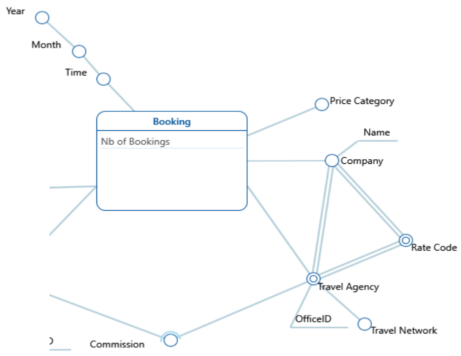
Algorithm 1 Cluster based popular ranking algorithm

RQUIRE: The priority of each dimensions and hierarchical structure of them. City and Company should be part of Dimensions

INPUT: `Dimensions`, a list contains the value of dimensions for aggregation . `K`, the number of hotels which we want to recommend

OUTPUT: `ListofHotels`, returns a list of most booked hotel in those context

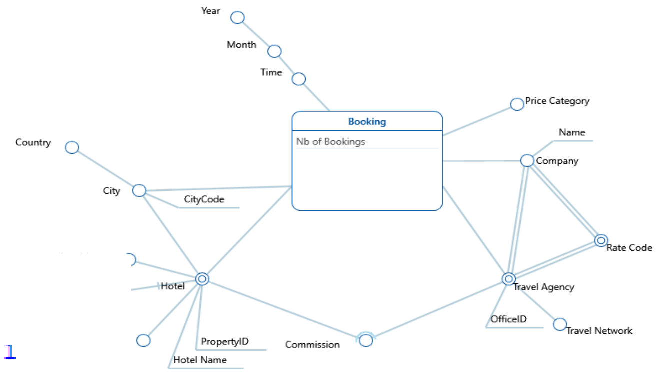
```
ListofHotels ← []
DictofHotels ← []
while DictofHotels != [] do
    DictofHotels ← [SELECT hotels, Count(*) groupby Dimensions]
    if DictofHotels == [] then
        if length(Dimensions) > 4 then
            Dimensions ← two dimensions with highest priority from Dimensions + [City,
            Company]
        else
            Dimensions ← [City, Company]
    else
        Dimensions ← [City]
DictofHotels ← Sort DictofHotels based on their Count
ListofHotels ← [SELECT Top K hotel from DictofHotels]
{If number of Input dimensions in addition of city and company are greater than 4,
we choose two other dimensions except city and company}
```



Introduction
Literature Review
Data preprocessing
Solution
Popularity base RS
Similarity Base RS
Matrix Factorization
Evaluation
Conclusion

Integration OLAP with RS

Cluster based popular ranking method



Property IDs for new company 'Samanu' in Paris

Company

Rate Code

City Code

Chain Cod

Country C

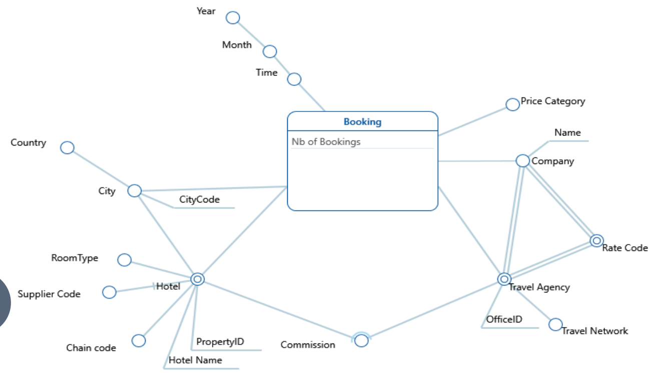
Year

```
RTPARDFS , MERCURE PARIS LA DEFENSE : 1901
SMPARDEF , MELIA PARIS LA DEFENSE : 1705
MDPARMER , LE MERIDIEN ETOILE : 1386
RTPARORL , NOVOTEL PARIS PORTE D'ORLEANS : 1211
RTPARORG , NOVOTEL POISSY ORGEVAL : 1095
RTPARLDF , NOVOTEL PARIS LA DEFENSE : 1080
RTPARARM , MERCURE PARIS GARE DE LYON TGV : 1026
RTPARLAD , IBIS PARIS LA DEFENSE CENTRE : 1002
BLPARP11 , BALLADINS ESBLY - MARNE-LA-VALLEE : 948
BLPARP12 , BALLADINS GENNEVILLIERS : 909
RTPAREXP , IBIS PARIS BERCY VILLAGE : 903
RTPARPLA , MERCURE PARIS PTE VERSAIL EXPO : 891
RTPARMTT , IBIS PARIS MONTMARTRE : 846
WIPAR729 , THE WESTIN PARIS VENDOME : 780
RTPARMTR , MERCURE MONTMARTRE SACRE COEUR : 756
AZPARDEF , CITADINES LA DEFENSE PARIS : 754
BLPARP09 , BALLADINS COIGNIERES : 740
PUPARARC , PULLMAN PARIS LA DEFENSE HOTEL : 691
RTPARMRG , MERCURE PARIS PTE D'ORLEANS : 679
RTPARISS , NOVOTEL SUITES PARIS ISSY : 677
RTPARLYO , NOVOTEL PARIS GARE DE LYON : 640
```

Submit

Integration OLAP with RS

Multi Dimensional Model



Similarity based method

a generalized grouping by on a coarse-grained level

Travel Network

City Code

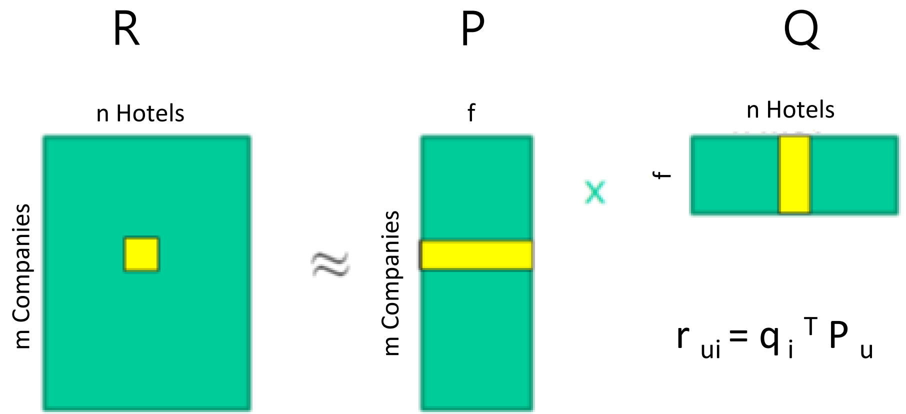
	Similarity	RMSE	MAE
SVD	Dot product	0.0541	0.0178
SVD++	Dot product	0.0002	0.0001
KNN	Pearson	0.0708	0.0291
KNN	Cosine	0.0589	0.0203

LBE

OSL

Matrix Factorization

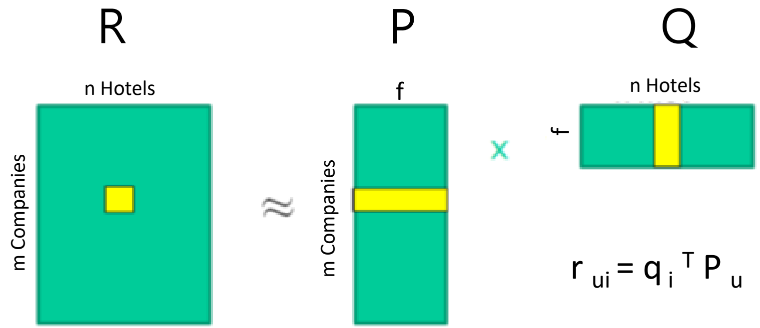
R = Rating Matrix, m Companies, n Hotels;
P = User Matrix, m Companies, f latent factors/features;
Q = Item Matrix, n Hotels, f latent factors/features;



$$r_{ui} = q_i^T P_u$$

A rating r_{ui} can be estimated by dot product of user vector p_u and item vector q_i

Matrix Factorization



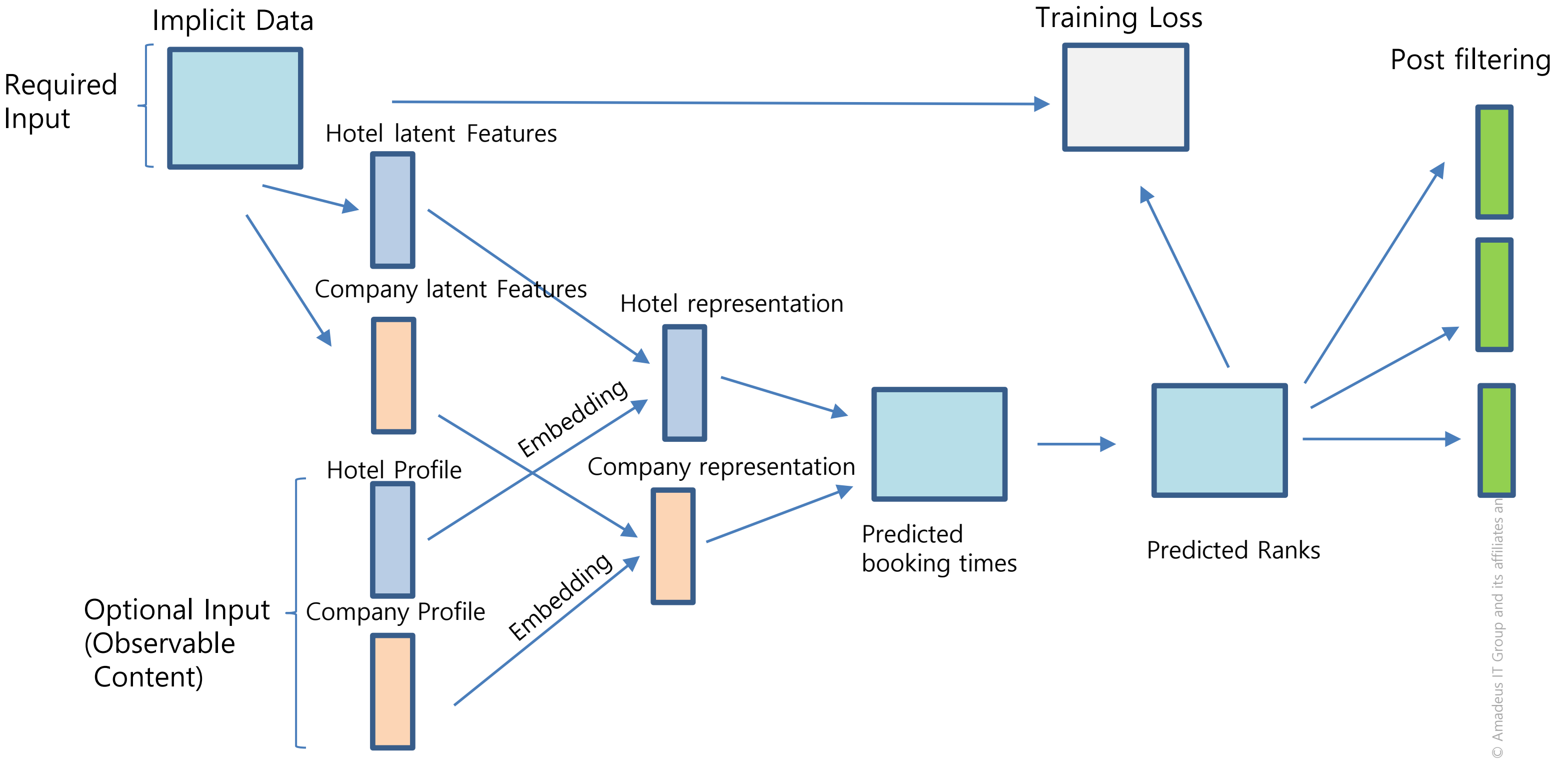
$$r_{ui} = q_i^T P_u$$

- p_u indicates how much user likes f latent factors;
- q_i means how much one item obtains f latent factors;
- The dot product indicates how much user likes item;
- The Latent Factor for companies could be “**Company Size**”, “**Industry Typ e** ”, “**Revenue Category**” and etc.

Matrix Factorization

- ❖ Fill in missing values
- ❖ Dimension Reduction
- ❖ Inferring unknown features (Unknown Business Policy)

Model


Implicit Data
Training Loss
Post filtering
Required Input
Hotel latent Features
Company latent Features
Hotel representation
Embedding
Company representation
Hotel Profile
Company Profile
Embedding
Predicted booking times
Predicted Ranks
Optional Input (Observable Content)

Company Profile

Top 4 cities

Average price

Top 4 most used rate codes

Months of Travelling

Number of Bookings

Average Length of Stay

Number of visited cities

Nordea Bank Danmark A/S

Banking company

Nordea Bank Danmark A/S is a bank in Denmark. It is part of Nordea - the largest Scandinavian financial group. Wikipedia

Headquarters: Copenhagen, Denmark

Number of employees: 31,596 (FTE, end 2016)

Traded as: Nasdaq Stockholm: NDA SEK; Nasdaq Helsinki: NDA1V

Founded: 1997

Key people: Björn Wahlroos (Chairman), Casper von Koskull (President and CEO)

Parent organization: Nordea

Nordea

Financial services company

Nordea Bank AB, commonly referred to as Nordea, is a Nordic financial services group operating in Northern Europe. Wikipedia

CEO: Casper von Koskull (Nov 1, 2015–)

Headquarters: Stockholm, Sweden

Revenue: 9.303 billion EUR (2016)

Total assets: 615.7 billion USD (2016)

Number of employees: 31,596 (FTE, end 2016)

Subsidiary: PlusGirot

Nordea Profile

Other features

```
{'STO': 1,
 'CPH': 1,
 'HEL': 1,
 'OSL': 1,
 'Nb_city_7': 1,
 'NRD': 1,
 'RAC': 1,
 '6R2': 1,
 'H4R': 1,
 'AvgPrice_2': 1,
 'Nbbooking_3': 1,
 'Month_march': 1,
 'Month_january': 1,
 'Month_december': 1,
 'Month_november': 1,
 'Month_october': 1,
 'Month_september': 1,
 'Month_feburary': 1,
 'NoNight_4': 1}
```

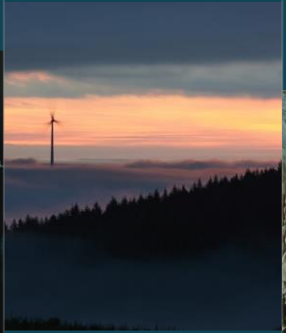
Average Price

Category 2:
8-27 Euro

Booking Times

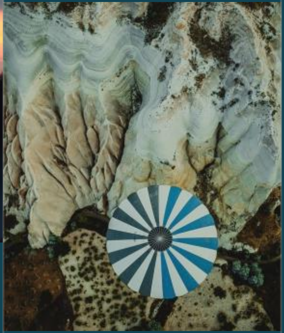
Category 3:
27-64

Months



March
January
December
November
October
September
February

Rate codes



NRD
RAC
6R2
H4R

Average length of stay

Category 4:
16-25 nights

Hotel Profile

City of the hotel

Average price

Top 4 most used rate codes

Supplier code

Months of Travelling

Number of Bookings

Average Length of Stay

Average Lead Time

Average Commission Price

Hotel Nice Riviera Profile

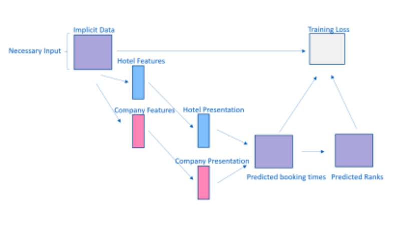
City	Average Price	Number of Booking	Average Commission	Supplier Code	Months	Top 4 rate codes	Length of Stay	Average Lead Time
Nice	Category 5: 64 - 125	Category 3: 8 - 27	Category 2: 1 – 2	Default	January October November September February	PR3 PR4	Category 2: 1 - 4	14 days in advance


```
{'city_nce': 1,
 'AvgPrice_5': 1,
 'Nbbooking_3': 1,
 'AvgCommission_2': 1,
 'suppliercode_def': 1,
 'Month_january': 1,
 'Month_october': 1,
 'Month_november': 1,
 'Month_september': 1,
 'Month_feburary': 1,
 'PR3': 1,
 'PR4': 1,
 'NoNight_2': 1,
 'AvgLeadTime_14': 1}
```

01

Matrix Factorization Without Embedding

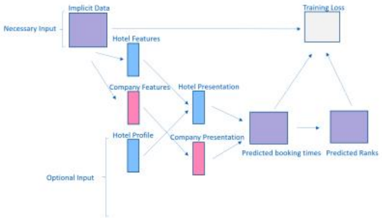
Considering Implicit data without considering side information.



02

Embedding Hotel Profile

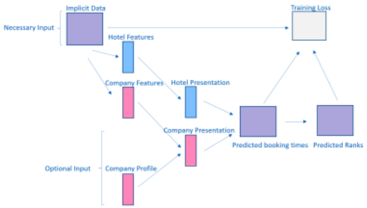
Considering Implicit data combined just with hotel profiles.



03

Embedding Company Profile

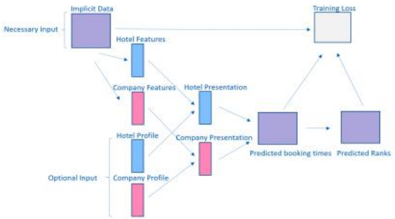
Considering Implicit data combined with company profiles.



04

Embedding both Profiles

Considering Implicit data combined with both profiles. Company and hotels



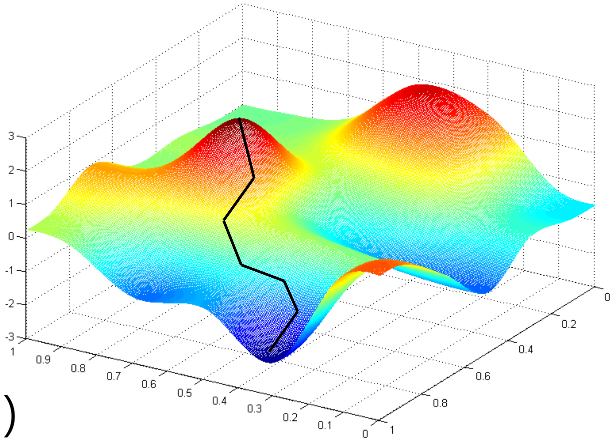
Learning Algorithms

❖ Stochastic gradient descent (SGD)

- Also known as incremental learning
- For each given training case, the system predicts r_{ui} and computes the associated prediction error

❖ Alternating Least Squares (ALS)

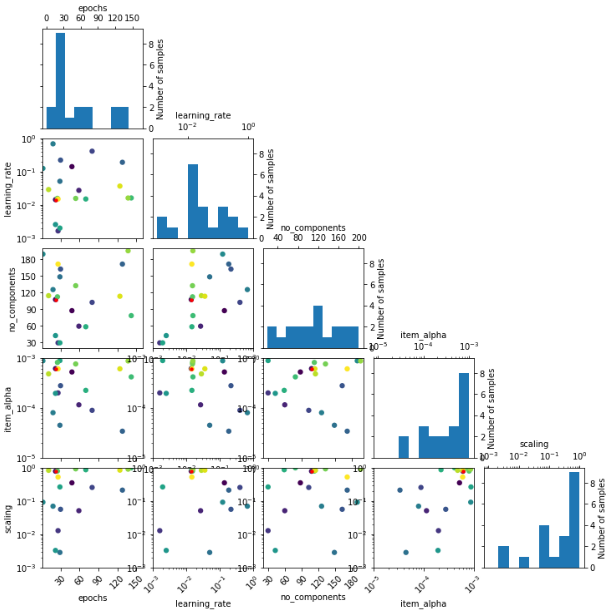
❖ Bayesian Personalized Ranking (BPR)

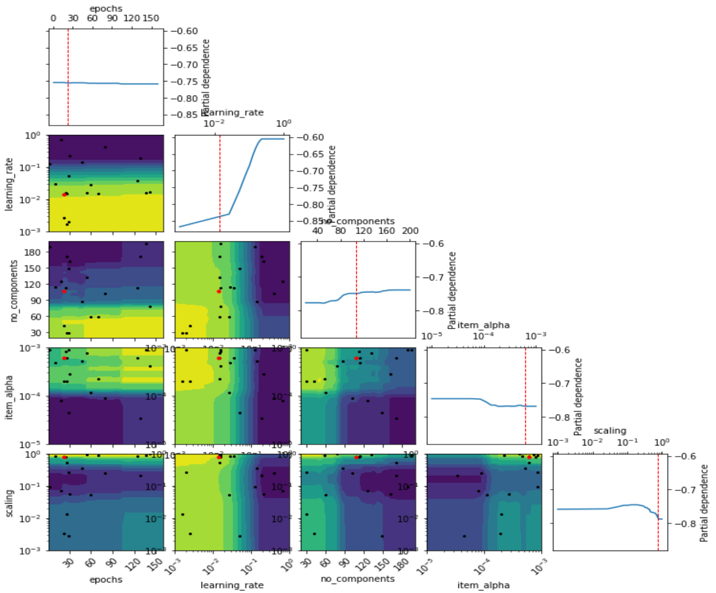


❖ Weighted Approximate-Rank Pairwise loss (WARP)

❖ Adaptive Moment Estimation (Adam)

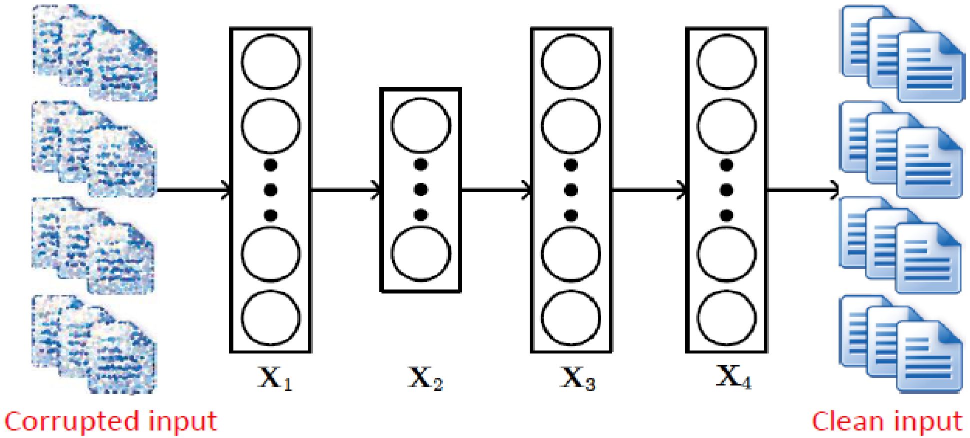
Hyperparameters





Model	Obj. value	epochs	learning_rate	no_factors	alpha	scaling
Model 1	0.91365	85	0.01845	106	3.184e-05	-
Model 2	0.92175	62	0.00950	103	1.856e-05	0.0697
Model 3	0.91330	22	0.01422	107	6.075e-04	0.77801
Model 4	0.92963	22	0.01422	107	6.075e-04	0.77801

Collaborative Deep Learning



Stacked Denoising Autoencoders

Yang, 2018

ALS	BPR	WARP	Adam	Popularity
WRMF	LightFM	LightFM	CDL	
Pointwise	Pairwise	Pairwise	Pairwise	Personalized popularity ranking per each user
Model_WRMF	Model_1_BPR Model_2_BPR Model_3_BPR Model_4_BPR	Model_1_WARP Model_2_WARP Model_3_WARP Model_4_WARP	PMF_OpenRec UserPMF_OpenRec	
1	4	4	2	1

Introduction

Literature Review

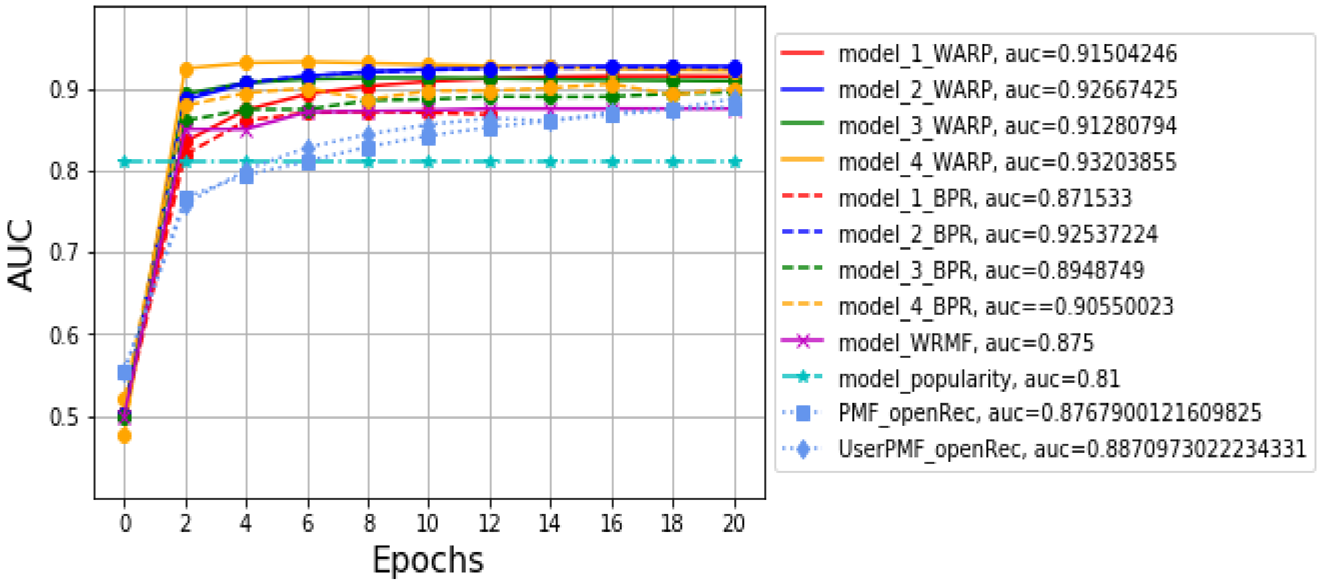
Data preprocessing

Solution

Evaluation

Conclusion

Evaluation



Introduction

Literature Review

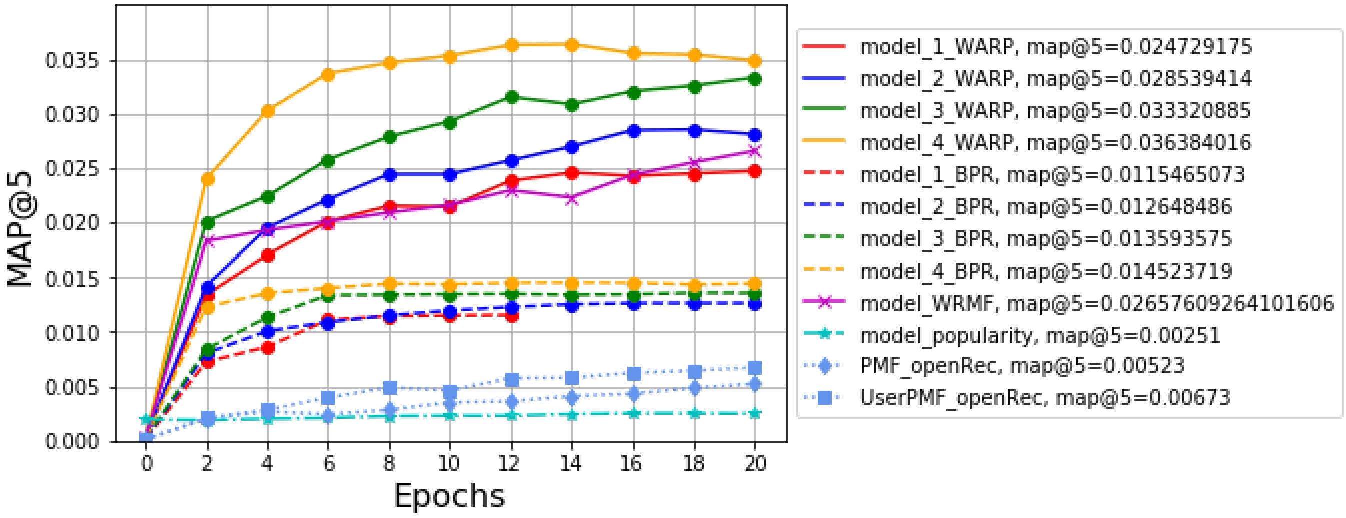
Data preprocessing

Solution

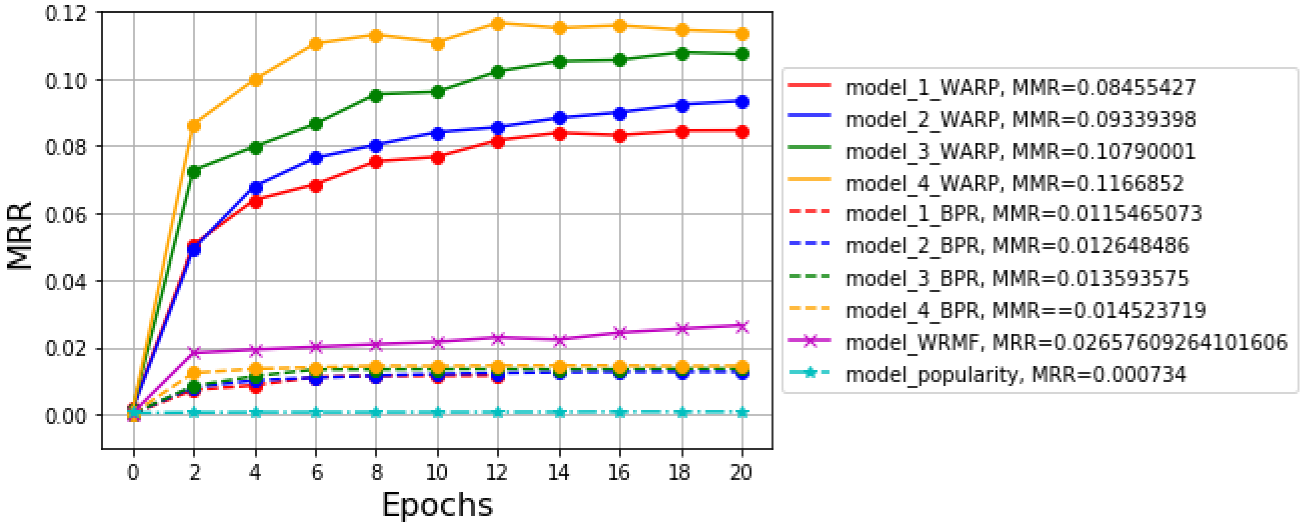
Evaluation

Conclusion

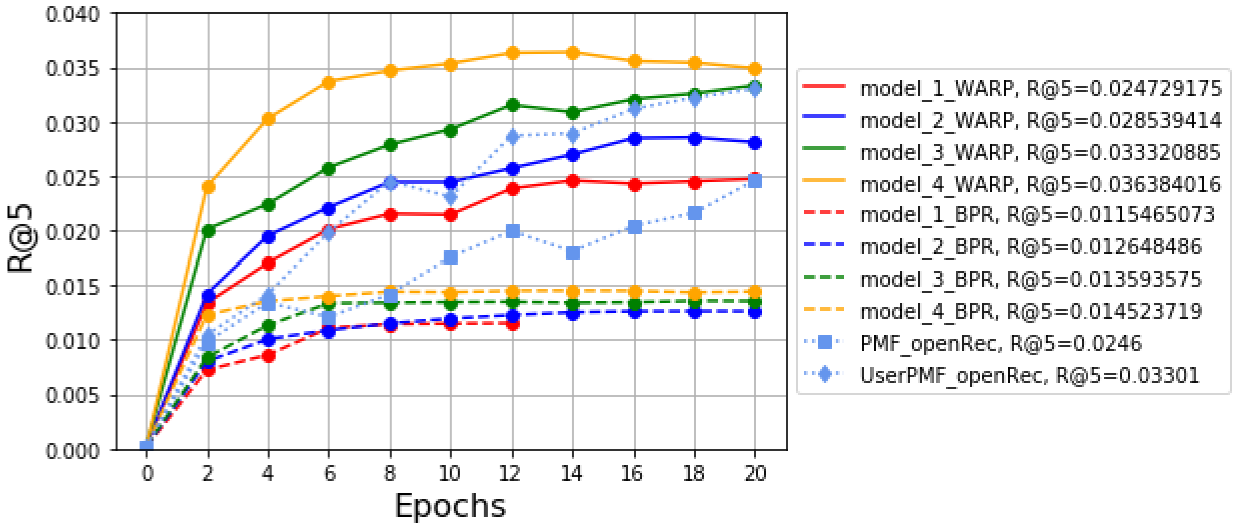
Evaluation



Evaluation

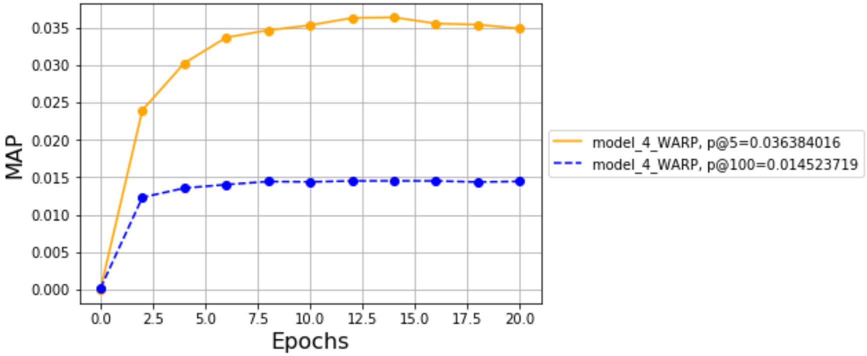

MRR
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
2
4
6
8
10
12
14
16
18
20
Epochs
model_1_WARP, MMR=0.08455427
model_2_WARP, MMR=0.09339398
model_3_WARP, MMR=0.10790001
model_4_WARP, MMR=0.1166852
model_1_BPR, MMR=0.0115465073
model_2_BPR, MMR=0.012648486
model_3_BPR, MMR=0.013593575
model_4_BPR, MMR==0.014523719
model_WRMF, MRR=0.02657609264101606
model_popularity, MRR=0.000734

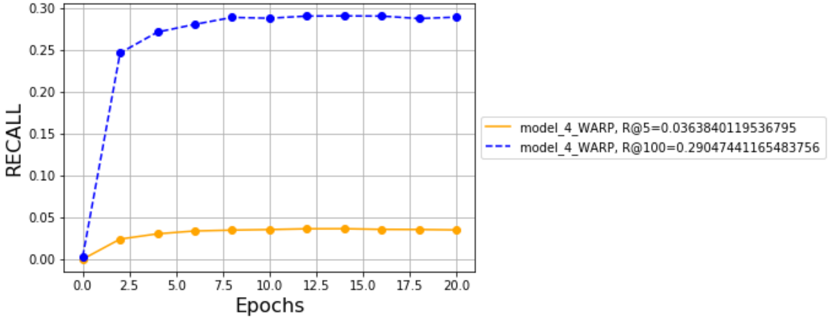
Evaluation


0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
R@5
0
2
4
6
8
10
12
14
16
18
20
Epochs
model_1_WARP, R@5=0.024729175
model_2_WARP, R@5=0.028539414
model_3_WARP, R@5=0.033320885
model_4_WARP, R@5=0.036384016
model_1_BPR, R@5=0.0115465073
model_2_BPR, R@5=0.012648486
model_3_BPR, R@5=0.013593575
model_4_BPR, R@5=0.014523719
PMF_openRec, R@5=0.0246
UserPMF_openRec, R@5=0.03301

Evaluation

Introduction
Literature Review
Data preprocessing
Solution
Evaluation
Conclusion





Introduction

Literature Review

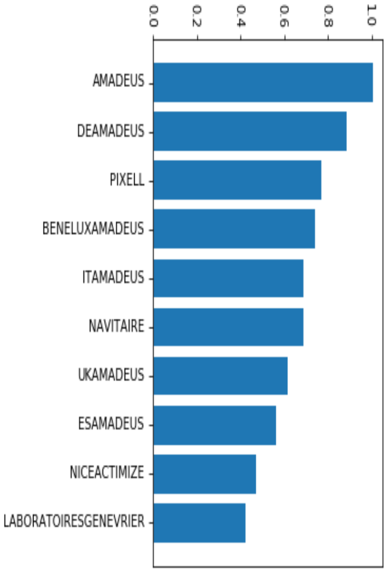
Data preprocessing

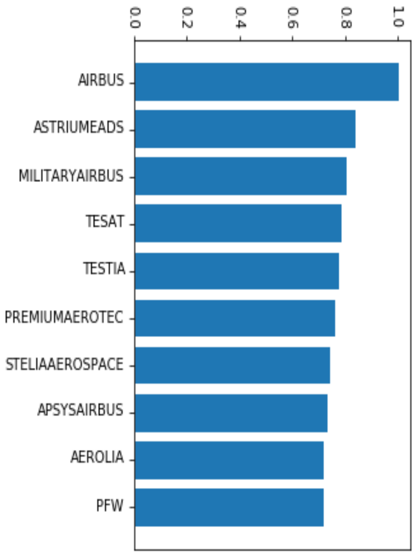
Solution

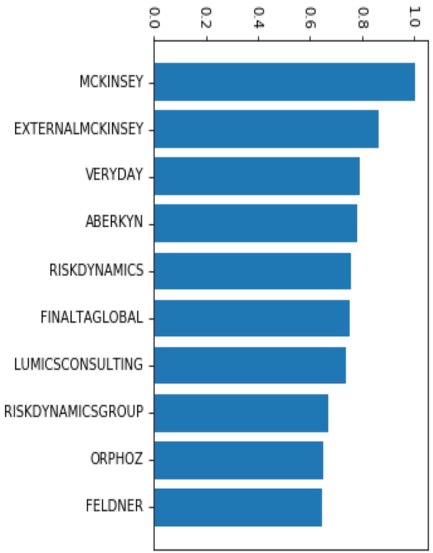
Evaluation

Conclusion

Evaluation







Post-filtering

City

OfficeID

Month

Sponsorship

...

f

w

c

b

y

Obligatory
This dimension is obligatory for our model

Separated RS
If we want different result for recommendations per each office ID, we can first filter based on the officeID and give more priority to the hotels were was booked already with the officeID.

Time
More accurate result based on the expected availability of the hotels in different months

Priority
Giving Priority to the hotels which they are sponsors

...
Any other kind of information can be Considered for post-filtering

Introduction

Literature Review

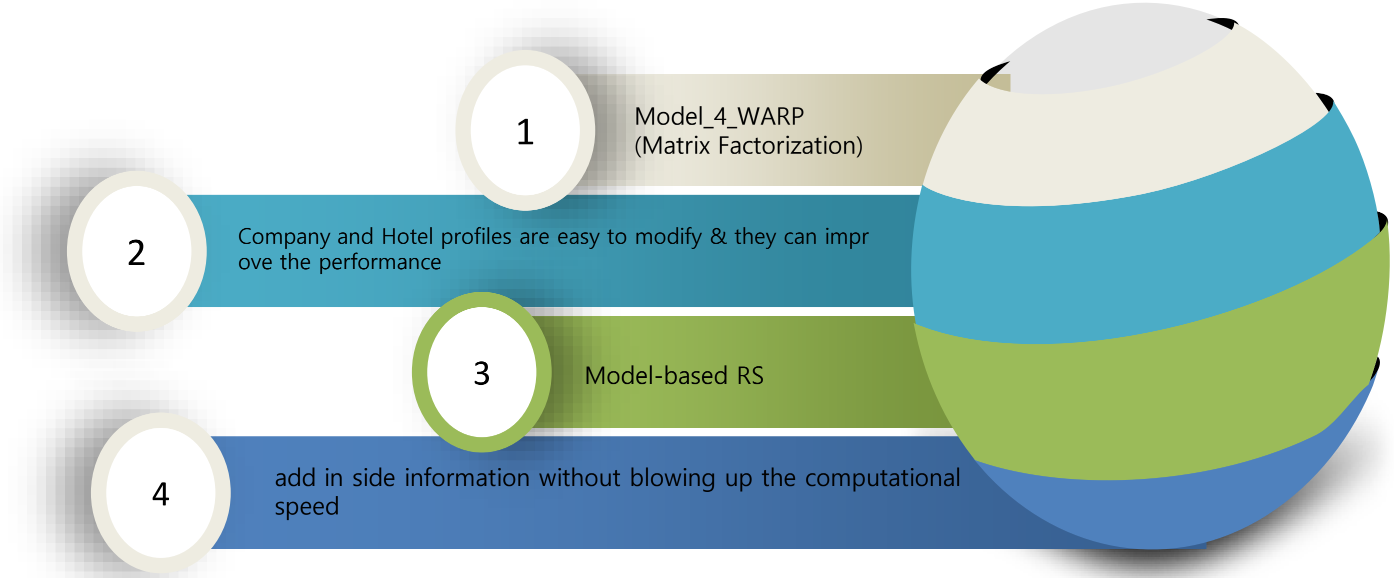
Data preprocessing

Solution

Evaluation

Conclusion

Conclusion



light fm

Python 3.6
16 GB RAM
Corei7

OpenRec

Future Work

1. Different Neural network model for same model
2. Experiment CDL method with all of raw data of companies and hotels
3. Factorization learning methods just with a few contexts
4. Adding sentiment or visual information as the input to our model