01
21-4-2021
Titus Pellegrom

Personalized Expedia Hotel Searches

Xinxing Jiang, Yao Xiao, and Shunji Li

Stanford University

02
Titus Pellegrom

December 13, 2013

Abstract

In this paper, we propose machine learning algorithms with search data of Expedia to solve personalized hotel ranking problem. First, we implement model selection and feature selection to improve classification result. We use different classifiers and implement backward search algorithm. In addition, we explored another alternative and implement collaborative filtering to rank hotels for each search item. Finally, we give conclusion and discuss our future work.

1 Introduction

The past decade of Internet evolution has led to the establishment of e-commerce markets. In the tourism industry, there is already a gradual shift towards electronic transactions. In recent years, growth rates in online tourism have moved much faster than in the overall world economy and travel expenses are the third largest cost-block in companies after labour and IT. Several prominant Online Travel Agencies (OTAs) are already offering personalized services to help custormers to choose their hotels, it is still the customers' job to take consideration of different aspects of hotels before final decision. Our goal is to apply machine learning techniques to rank hotels, which is based on a particular customers' characteristics and recommend those with high matching scores specifically to him/her.

3-4
2 notes:

05
Titus Pellegrom

Some hotel selection and recommendation engines are based on Knowledge based systems and Knowledge Modelling [1]. Moreover, techniques such as Semantic Web-based extended the matching process to improve the precision. Collaborative filtering, used as one of the most effective method of dealing with recommender problems, are also studied. A hotel recommendation system based on collaborative filtering method of clustering and Rankboost algorithm has been proposed[3]. This is also the basis for our final algrithm discussed in this paper.

The paper however, elaborates the procedures of finding a better algorithm for the construction of a learning model for personalized hotel selection and recommendation. Our data, which includes features such as hotel characteristics, hotel location, user's aggregate purchase history and competitive Online Travel Agency's information, will be used to design a machine learning algorithm to rank these hotels in such a way that hotels with higher rankings are those with a higher probability to be clicked/booked. Our work will include model selection, optimization and feature selection.

This paper is organized as follows. Section 2 describes the data sets we use. Section 3 shows the model selection process. Section 4 shows the feature selection. Section 5 shows the collaborative filtering we use as an alternative to rank hotels. Finally, section 6 gives our conclusion and discusses the future plan of our work.

2 Dataset

The data provided by Expedia consists of a list of hotel search queries, and each row of a query is associated with a particular hotel. After data cleaning, our training set contains about 10 million data items. Each item contains over 50 features. Since the dataset is huge and the data items are randomly organized, we will not consider trying to train our algorithm on a small subset of the whole dataset. And since some features (date and time) are obviously irrelevant to our prediction result, we will not consider them during our training process.

3 Model Selection

Due to the nature of consumer's online activity, both the click through rate and the ac- tual booking rate are extremely low. (In our first 10000 training set, the click through rate is 0.0457 and the booking rate is about 0.0281). Therefore, we did not apply the classification error to determine the efficiency of our training algorithm but Instead, we chose the precision and recall metric (fscore) to evaluate the performance of our training algorithm.

During the process of trying to fit our data to a better model, we applied several common machine learning classifiers but neither precision nor recall rate can reach a desirable level. The SVM and logistic regression algorithm always predict the click through/booking to be false and moreover, their precision and recall rate fails to make significant improvements with the growing size of training sets. The best precision and recall rate we could achieve after applying perceptron and backward feature selection were 0.96 and 0.38, which is still a little disappointing. Therefore, we came up with a hypothesis that Collaborative Filtering could work better for this problem and more details will be discussed in the Future Work section.

4 Feature Selection

We implement feature selection in the following two ways.

4.1 Preprocess

We manually remove some features which are obviously irrelevant to our prediction to reduce noise and improve time efficiency.

4.2 Backward Search

We implement backward search algorithm, and use k-fold cross validation to choose best features from the rest features. In particular, for each feature, we compute F-Score of two cases, say removed or preserved, and choose a better case.

Each classifier gives different evaluation of same feature set. For Perceptron classifier, the feature set finally removed is:

{prop_country_id, prop_review_score, srch_length_of_stay, srch_children_count, srch_query_affinity_score, comp1_inv, comp1_rate_percent_diff, comp2_inv, comp3_inv, comp3_rate_percent_diff, comp4_rate, comp4_inv, comp5_rate_percent_diff, comp6_rate, comp6_inv, comp6_rate_percent_diff.}

After feature selection, we can achieve F-Score of 0.96 (take not book as true) and 0.38 (take book as true).

5 Collaborative Filtering

In this part, we choose another method to rank hotels based on collaborative filtering, a model used in recommendation systems. We notice a phenomenon in real life that people who click or book similar hotels should be similar. In order to make use of this phenomenon, we implement collabrative filtering to rank hotels for a specific search as shown in the following:

(a) We calculate the average of features of all the queries that click through or book the hotel for each of the hotels and let it be the hotels profile feature.

(b) For each (search id, hotel id) pair, we calculate the distance between the querys features and the hotels profile based on several similarity metrics.

(c) For each unique search id, we rank the hotels from low to high according to the numerical value of the distances calculated in the previous step.

(d) We evaluate quality of the rank based on a score related to weight and rank and choose the best similarity metric.

In order to reduce the risk of unbalanced feature range, we compute the mean value and standard derivation value of each hotel to normalize features in each search item.

Figure 1 shows the evaluation of different similarity metrics. The third column of Figure 1 shows the percent of score of the rank and total score. The total score is computed in the following way:

(a) If a hotel is clicked, we add 1 point to the total score.

(b) If a hotel is booked, we add 5 points to the total score.

Table 1: Evaluation of Different Simialrity Metrics

Similarity Metric	Score	Percent (%)
Hamilton	255831.379231	56.17
Euclidean	271733.655565	59.66
Third Order	275474.194899	60.48
Cosine	315820.196249	69.34

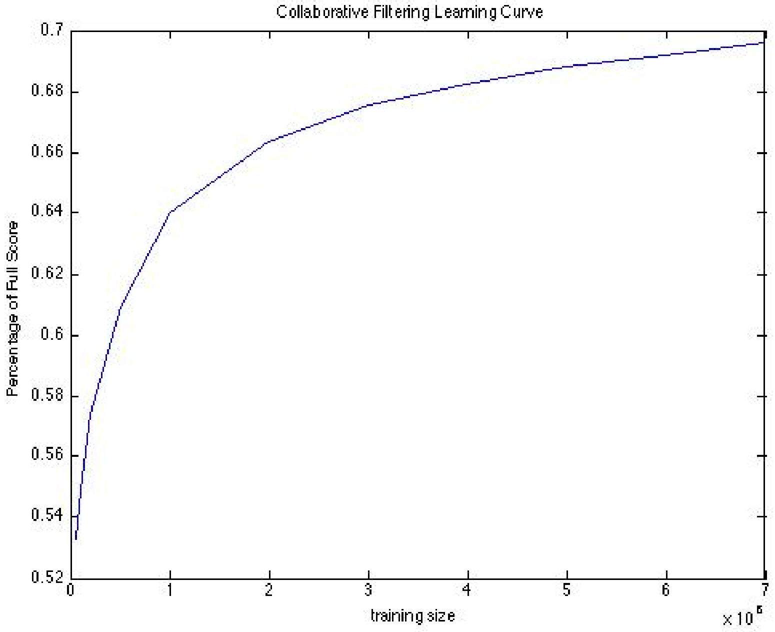
To convert from a rank to a score, we use the following fomula.

$$Weight = 2 - \frac{2}{1 + e^{0.1(-rank+1)}} \tag{1}$$

We discover the number of hotels corresponding to a search item is about 30. And we need a convert formula to map rank to score. In specific, we need to map rank No.1 to 1, and rank No.30 to about 0. After several trails, we choose the above formula.

We run the baseline algorithm which uses ramdom forest to rank the hotel. According to previous metric, the rank can achieve 50% of total score. In addition, baseline algorithm needs several hours to run, while our algorithm runs much faster, say less than one minute.

Below is a plot of our learning curve, the y axis shows the percentage of our score as compared to the maximum score we can achieve:



6 Conclusion and Future Work

In this paper, we implement personalized hotel ranking based on collaborative filtering. We analyze the phenomenon of similarity between people who click or book similar hotels, and design our algorithm based on the finding. Compared to baseline algorithm, our algorithm is much faster and is online. Our work raises a number of questions about how to deal with sparse profiles (which contains many NULL values) of the search data. Currently, we simply replace NULL values with 0, which is not ideal. We might use mean values or ramdom sampling to deal with NULL values. Another main focus of our future work is to combine and modify different models instead of using just one classifier. Limited by time and computation resources, more detailed analysis will be conducted in the future.

7 Acknowledgement

We express our sincere thanks to Kaggle and Expedia for offering us the dataset.

References

[1] B. A. Gobin, and R. K. Subramanian *Knowledge Modelling for a Hotel Recommendation System.* World Academy of Science, Engineering and Technology 1 2007

[2] T. Berners-Lee, J. Hendler, and O. Lassila *The Semantic Web.* Scientific American, pages 3443, May 01.

[3] Gao Huming, Li Weili *A Hotel Recommendation System Based on Collaborative Filtering and Rankboost Algorithm.* 2010 Second International Conference on MultiMedia and Information Technology(MMIT), pages 317-320, April 2010

Personalized Expedia Hotel Searches

Zhang, Owen

01 Titus Pellegrom Page 1

21/4/2021 13:47

Waardeloze paper

02 Titus Pellegrom Page 1

21/4/2021 13:47

03 Titus Pellegrom Page 1

21/4/2021 13:41

04 Titus Pellegrom Page 1

21/4/2021 13:41

05 Titus Pellegrom Page 1

21/4/2021 13:41

06 Titus Pellegrom Page 2

21/4/2021 13:44

Why not?

07 Titus Pellegrom Page 2

21/4/2021 13:44

08 Titus Pellegrom Page 2

21/4/2021 13:44

Geldt niet voor DMT data set!

09 Titus Pellegrom Page 2

21/4/2021 13:44

10 Titus Pellegrom Page 5

21/4/2021 13:46

??

11 Titus Pellegrom Page 5

21/4/2021 13:46