ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 150-155
doi: 10.18178/wcse.2019.06.022

Latent Factor-based Rating Feedback Learning for Restaurants Recommendation

Yi Xu, Ziliang Wan, Zige Zhou, Yuchen Liu and Jinpeng Chen[+]

School of Software Engineering, Beijing University of Posts and Telecommunications, China

Abstract. Nowadays, when people go out to eat, their choice of restaurant depends not only on taste, but also on many other factors. Therefore, mining what factors of the restaurant the users care about is a key problem for restaurant recommendation. This paper is engaged to mining the latent theme factors of restaurant the users care about and applying the result to restaurant recommendation. In this paper, we used LDA model to extract the latent theme features of the restaurants, calculated the similarity based on latent factors and rating feedbacks to make rating prediction and restaurants recommendation. This paper conducted an experiment with the review data from Yelp dataset, exploring the performance of the algorithm and the optimal theme number K. The results of the experiment showed that the algorithm achieved some improvement in rating prediction. To some content, applying the latent theme distribution to the problem of restaurant recommendation can solve the problem of data sparsity, decrease the computational dimension and raise the accuracy of rating prediction.

Keywords: Latent Factor; Rating Feedback; Restaurants recommendation; LDA

1. Introduction

In recent years, the catering industry has developed rapidly, and many restaurants have made innovative attempts in dishes, dining services and restaurant decoration styles. The factors young people consider in choosing restaurants have also become diversified, not only the taste of food. Some social medias of restaurants have provided a multi-aspect rating mechanism, for example, Dianping.com provides an environment-taste-service three aspect rating service. Under such rating, the user rating information is still very weak for reflecting restaurant characteristics. There is a need for multi-dimension rating information containing the diverse factors and aspects to represent the characteristics of the restaurants.

Food experts can make a detailed classification of dishes, but it is difficult to control the granularity of classification and set the weight of a restaurant in a certain category. The item-CF makes recommendations based on users' historical behaviors and gives corresponding explanations. These explanations usually make users more convinced. This paper is attempted to build a model considering the latent, diverse factors of restaurants and the customers' historical behaviors to make effective recommendations for customers.

Previous studies show that user-provided reviews include richer information than ratings in the process of recommender[1], as a result, this paper is engaged in mining diverse latent theme factors and aspects of restaurants from users-generated reviews.

This paper proposed an LDA-based collaborative filtering algorithm which takes account of the latent theme factor mining from reviews and the customers' historical rating feedbacks, and carried out experiments to verify the effectiveness of the algorithm. Our contributions are:1) recommending restaurants to users not just taking account of taste factor but diverse latent theme factors;2) extracting the latent theme distribution features of restaurants and form the latent K-theme ratings matrix to analysis users' preferences; 3) a method to apply the diverse latent theme factors to restaurants recommendations.

[+] Corresponding author. Tel.: + 86 13811415176 ; fax: +86 58828012
E-mail address: jpchen@bupt.edu.cn.

This article consists of eight parts. The first part introduced the origin and background of the problem and the second part introduced related work. The third part made the definition of the problem, the fourth part described the model our proposed. The fifth part described the experiment we carried out and analyzed the results of the experiment. The sixth part described the conclusion as well as the next step of work. The seventh part is our acknowledgment. The last part listed all the references.

2. Related Work

Lihua Sun et al. applied uncertainty theory into the restaurant recommender system based on sentiment analysis of online Chinese reviews to mining users' opinions[1]. Sonya Zhang et al. exploited a recommender system for restaurant reviews based on consumer segment, proposing a content-filtering recommender system that evaluates individual online reviews and assigns a numeric score to each review for each of the five consumer segments[2]. These papers inspired this article. In order to mine diverse latent theme aspects and features, this paper can make a restaurant feature segment based on mining the users' reviews.

Chao Li, Srn Feng et al. applied a LDA and WordNet combined algorithm to mining dynamics of research topics and got kind of great results[3]. Maha Amami et al. applied LDA-Based Approach to scientific paper recommendation[4]. Shinjee Pyo et al. used LDA models to analysis the TV user groups and TV program, group similar TV users and associate description words for watched TV programs at the same time in a unified topic modeling framework[5]. LDA model shows great superiority in latent theme features and indicating the proper category.

Previous studies have suggested that restaurant customer reviews can be categorized into multi-factors such as service quality, product quality, menu diversity, price and value, atmosphere, etc[2]. Yifan Gao et al. built a restaurant recommender system based on a novel model that captures correlations between hidden aspects in reviews and numeric ratings and demonstrated the advantages by experiments[6]. Therefore, it is worthwhile and feasible building a restaurant recommender system based on LDA model and item-CF which taking account of latent factors mining from reviews and the customers' historical behaviors.

3. Problem Definition

Formally, there is a set of users U (for each user u, $u \in U$), and a set of restaurants I(for each restaurant i,$i \in I$). For each restaurant i, it has its own reviews set R_i, $R_i = \left\{r_{(u_j,i)}, reviews\ associated\ with\ restaurant\ i\ from\ user\ u_j,\ \forall\ u_j \in U\right\}$.The problem is to predict the ratings of users to the restaurants which they haven't been to and recommend high rated restaurants.

To solve this problem, this paper attempts to predict the ratings based on a modified item-CF. In traditional item-CF, first of all, calculate the similarity between items based on the user-item rating matrix. Then, select the first K items with the largest similarity with the target item to construct the neighbor set. Finally, predict the unknown ratings of target items by the weighted sum of known ratings of neighbor set. In this paper, we seek to exploit the latent factor from the reviews to calculate the similarity between items rather than abstract features from the user-item rating matrix.

4. A LDA-based Recommendation Model

In this section, this paper firstly introduced the latent theme distribution features by mining the reviews through the LDA topic model. Then, we introduce the calculation of the similarity. Finally, we introduce the ratings prediction and generating recommendations.

4.1. Latent theme distribution features

This paper attempts to construct the latent theme distribution features which contain the latent factors the users care about to represent the characteristics of restaurants. Specifically, in order to find latent and user-cared factors, this paper is engaged to find the themes frequently commented by users and infer how often the theme is commented in the reviews of the restaurant. With this goal, this paper adopts the LDA topic model to obtain the latent theme distribution from the reviews of restaurants.

Formally, to abstract the latent theme distribution features of the restaurant, this paper merged the reviews of one restaurant into a single document, which means merging reviews in set R_i (the set of reviews associated with restaurant i) into document d_i. Then, for the restaurants in set I, we build a set $D = \{d_i, \forall\, i \in I\}$. The idea of the Latent Dirichlet Allocation(LDA) topic model is showed in figure 1.

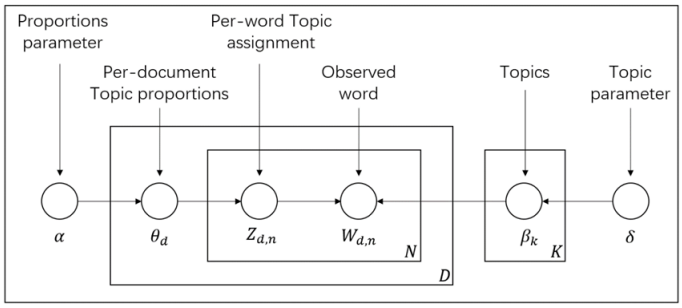


Figure 1: the Latent Dirichlet Allocation(LDA) topic model

For any document d, its topic distribution is θ_d,$\theta_d = Dirichlet(\vec{\alpha})$, where, α is the hyperparameter of the distribution and is a K-dimensional vector. K is the number of topics. For any topic k, its word distribution is β_k,$\beta_k = Dirichlet(\vec{\delta})$, where, δ is the hyperparameter of the distribution and is a V-dimensional vector. V is the number of words. For the n_{th} word of any document d, we can get the distribution of its topic number $Z_{d,n}$ from θ_d as $Z_{d,n} = multi(\theta_d)$, for the topic number $Z_{d,n}$, we can get the distribution of its words $W_{d,n} = multi(\beta_{Z_{d,n}})$. Given the prior parameters α and δ, the joint probability distribution of LDA is $P(\theta, \beta, Z|W, \alpha, \delta) = \frac{P(\theta,\beta,Z,W|\alpha,\delta)}{P(W|\alpha,\delta)}$. There are many methods to estimate parameters of LDA model, including Variational Bayes Inference, Gibbs Sampling and Maximum Likelihood Estimation.

In this paper, we input the document d_i for each restaurant i and set the number of topics as K, then we use Variational Bayes Inference to get the parameters. This paper used the document-topic distribution θ_{d_i} as the latent theme distribution feature of restaurant i.

4.2. Similarity calculation

The latent theme distribution feature reflects the themes the users care about and the intensity of the user's focus on this theme, but can't reflect whether the theme is good or bad. Therefore, we decompose the ratings into the latent theme distribution, obtaining the item-K themes ratings matrix. To some extent, the item-K themes ratings matrix is almost identical to the user's detailed rating information for various aspects of the restaurant.

Use the item-K themes ratings matrix to calculate the cosine similarity between restaurants. For example, suppose the latent theme distribution features of restaurant m and n are vector $\overrightarrow{\vartheta_m}$ and $\overrightarrow{\vartheta_n}$, the ratings of restaurant m and n are r_m and r_n. Thus, the similarity between restaurant m and n is $\text{sim}(m,n) = \cos(r_m \cdot \overrightarrow{\vartheta_m}, r_n \cdot \overrightarrow{\vartheta_n}) = \frac{(r_m \cdot \vartheta_m)\cdot(r_n \cdot \vartheta_n)}{||r_m \cdot \vartheta_m|| \cdot ||r_n \cdot \vartheta_n||}$.

4.3. Ratings prediction and recommendations

After calculated the similarity between restaurants using the item-K themes ratings matrix, use the item-based collaborative filtering algorithm to predict the ratings and get Top-K restaurants to form the recommended list. For user u and restaurant i, the format of the rating prediction is as follows:

Equation 1: Rating prediction

$$\widehat{r_{ui}} = \frac{\sum_{j \in N_u^k(i)} sim(i,j) \cdot r_{uj}}{\sum_{j \in N_u^k(i)} sim(i,j)}$$

$N_u^k(i)$ *is the neighbor set of restaurant* i *whrn the topic number is* K.

5. Experiment

5.1. Dataset

This paper experimented on part of the Yelp Dataset which was provided by Yelp Dataset Challenge Round 12. The whole dataset included information about 188 thousand local businesses in 10 metropolitan and 5996995 reviews from 1518169 users to the 188593 businesses. Here we conducted the experiment with

the first 25176 comment records from the review data, which contains the ratings and corresponding comments of 18,479 restaurants from 10,000 users.

Firstly, we merge reviews into documents by restaurants. Secondly, preprocess the documents using the nltk library. Then, build LDA model for documents using the sklearn library to get the latent theme distribution of restaurants and decompose the ratings to K-themes to compute the similarity between restaurants. Finally, we apply the similarity to the collaborative filtering algorithm for scoring prediction and Top-K recommendation.

This experiment uses MAE(Mean Absolute Deviation, see Equation 2)，RMSE(Root Mean Squared Error, see Equation 3)，FCP(Fraction of Concordant Pairs) the to measure the accuracy of the rating prediction of the algorithm under 5-folds cross validation（ 80% of the dataset we used as the trainset, and 20% as the testset）. MAE reflects the absolute error level of the rating prediction algorithm, RMSE reflects the stability of accurately predicting ratings, while FCP means the proportion of concordant pairs between predicting results with the actual data.

Equation 2: MAE

$$\mathrm{MAE} = \frac{1}{|\hat{R}|}\sum_{\widehat{r_{vi}}\in\hat{R}} |r_{vi} - \widehat{r_{vi}}|$$

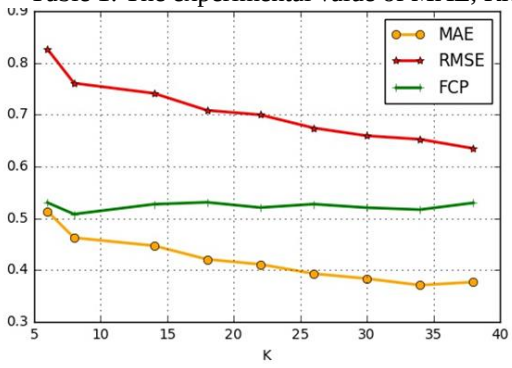
Equation 3: RMSE

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\hat{R}|}\sum_{\widehat{r_{vi}}\in\hat{R}} (r_{vi} - \widehat{r_{vi}})^2}$$

5.2. The optimal topic number K

In this paper, the calculation of similarity depends on the latent theme distribution extracted from the LDA model, the topic number K has a certain impact on the accuracy of the algorithm. Therefore, this experiment explored the optimal topic number K, conducting the experiment for K∈[6,38] and K∈N with a step of 4. The experimental results are shown in the figure and table below. (see Figure 2 and table 1)

Table 1: The experimental value of MAE, RMSE, FCP searching for the optimal theme number K



K	MAE	RMSE	FCP
6	0.5135	0.8266	0.5303
10	0.4625	0.7609	0.5080
14	0.4471	0.7413	0.5225
18	0.4208	0.7084	0.5309
22	0.4111	0.6998	0.5207
26	0.3931	0.6744	0.5274
30	0.3836	0.6593	0.5205
34	0.3709	0.6525	0.5167
38	0.3769	0.6349	0.5298

Figure 2:The experimental value of MAE, RMSE, FCP searching for the optimal theme number K

As shown in the figure, when the value of K increases, the value of MAE and RMSE decreases, which means the accuracy of the algorithm is enhanced and the stability of the algorithm performance grows stronger. The value of FCP fluctuates in a very small range as the K increases, which means there is no significant increase in the number of concordant pairs. As a result, the variety of K has little effect on improving the level of concordance of accurately rating prediction. On the whole, the increase of K can significantly reduce MAE and RMSE, however, without significantly affection on the change of FCP, which means the accuracy and stability of individual score prediction would be enhanced, but the overall prediction concordance wouldn't be greatly improved.

On the other hand, when the value of K increases, the interpretability of these topics decreases and the computational complexity of the algorithm increases. Based on the above analysis, the optimal topic number K is 38 for K∈[6,38] and K∈N with a step of 4.

5.3. Algorithm performance

Under the same conditions, the experiment was carried out on the same data using the traditional item-based collaborative filtering algorithm. The result is: MAE=0.6083, RMSE=0.9334, FCP=0.4765. Compared

with the traditional collaborative filtering algorithm, generally, the MAE and RMSE of this algorithm are smaller, so, the level of accuracy and stability of this algorithm is enhanced. FCP is larger in this model, as a result, the concordance is improved in this algorithm(see Table 2). With this algorithm, the accuracy and stability of individual score prediction would be enhanced, and the overall prediction concordance would be improved in some content. We can conclude that LDA topic model performs kind of good performance in restaurant feature extraction and restaurant recommendation as well as text-related recommendation

Table 2:Comparing the MAE, RMSE, FCP of this model with traditional CF

K	MAE	MAE Reduced	RMSE	RMSE Reduced	FCP	FCP Increased
6	0.5135	0.0948	0.8266	0.1068	0.5303	0.0538
10	0.4625	0.1458	0.7609	0.1725	0.5080	0.0315
14	0.4471	0.1612	0.7413	0.1921	0.5225	0.0460
18	0.4208	0.1875	0.7084	0.2250	0.5309	0.0544
22	0.4111	0.1972	0.6998	0.2336	0.5207	0.0442
26	0.3931	0.2152	0.6744	0.2590	0.5274	0.0509
30	0.3836	0.2247	0.6593	0.2741	0.5205	0.0440
34	0.3709	0.2374	0.6525	0.2809	0.5167	0.0402
38	0.3769	0.2314	0.6349	0.2985	0.5298	0.0533

5.4. Results and discussion

This paper used the item-K themes ratings matrix to compute the similarity, to some content, solving the problem of data sparsity caused by the sparse user-item rating matrix and reducing the computational dimensions from the number of users to the topic number. The item-K themes ratings matrix described the features of the restaurant from K aspects, which is more detailed than the user rating matrix, thus bringing more accurate similarity and rating prediction.

6. Conclusion and Future Work

This paper studied the restaurant recommender system from the perspective of mining diverse latent themes the users care about from reviews. We designed the latent theme distribution feature and the item-K themes ratings matrix to integrate the characteristics contained in the latent factors and rating records, and apply them to the similarity computation. This paper also conducted an experiment to explore the optimal theme number K and compare the performance with the baseline algorithm--Item CF. According to the results, we can conclude that mining the latent aspects the users care about can be helpful to tackle the restaurant recommend problem.

In the future, we plan to conduct emotional analysis on the word distribution of K themes extracted from the LDA model, and then predict the ratings of K aspects based on the results of emotional analysis to improve the accuracy of the algorithm.

7. Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No.61702043, and Research Innovation Fund for College Students of Beijing University of Posts and Telecommunications.

8. References

[1] Lihua Sun, Junpeng Guo, Yanlin Zhu:Applying uncertainty theory into the restaurant recommender system based on sentiment analysis of online Chinese reviews. World Wide Web 22(1): 83-100 (2019)

[2] Sonya Zhang, Mohammad Salehan, Andrew Leung, Ishmene Cabral, Navid Aghakhani. A Recommender System for Cultural Restaurants Based on Review Factors and Review Sentiment. AMCIS 2018

[3] Chao Li, Sen Feng, Qingtian Zeng, Weijian Ni, Hua Zhao, Hua Duan:Mining Dynamics of Research Topics Based on the Combined LDA and WordNet. IEEE Access 7: 6386-6399 (2019)

[4] Maha Amami, Gabriella Pasi, Fabio Stella, Rim Faiz. An LDA-Based Approach to Scientific Paper Recommendation. NLDB 2016: 200-210

[5] Shinjee Pyo, Eunhui Kim, Munchurl Kim. LDA-Based Unified Topic Modeling for Similar TV User Grouping and TV Program Recommendation. IEEE Trans. Cybernetics 45(8): 1476-1490 (2015)

[6] Yifan Gao, Wenzhe Yu, Pingfu Chao, Rong Zhang, Aoying Zhou, Xiaoyan Yang:A Restaurant Recommendation System by Analyzing Ratings and Aspects in Reviews. DASFAA (2) 2015: 526-530

[7] Koren Y , Sill J . Collaborative filtering on ordinal user feedback[C] Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. 2013.