Research on location and similar comments in point-of-interest recommendation system for users

Yu Liu¹,², Ning Wei¹

¹College of Mathematics and Information Science, Hebei University, Baoding, Hebei, 071002
²Hebei Key Laboratory of Machine Learning and Computational Intelligence

Corresponding author and email: Ning Wei, zhangming@hbu.edu.cn

Abstract. By combining deep learning algorithms and potential probabilistic topic models, this paper proposes a new modeling strategy for individual users and constructs an effective point of interest recommendation system. First, Doc2vec is used to vectorize the user's comments, and the Bow model is used to match the similarity of the points of interest, and then the potential probability model is used to cluster the positions of the points of interest. Finally, relying on the user's historical comment data as a basis, the Top-K ranking recommendation scheme is obtained. The experimental results show that for large-scale location-based social network data sets, the algorithm in this paper has an ideal final effect on users' personalized recommendations. The recommendation model for user location information and similar comment points of interest proposed in this paper is more subjectively and objectively in line with user behavior needs. On the premise of ensuring the real-time performance of recommendation, the model proposed in this paper effectively improves the accuracy of recommendation compared with traditional point-of-interest recommendation algorithms.

1. Introduction

1.1. Developing status

The personalized recommendation system recommends information and products of interest to users based on the user's points of interest and historical purchase behavior.

Literature [1] shows the influence of data sparsity on the experimental results. In the literature [2], the lack of data also has a negative impact on the experimental results. Literature [3] considers the relevant information contained in the comments and associates the comments with the content posted by users. Finally, through collaborative filtering for recommendation, a new recommendation model is proposed to improve the accuracy of recommendation. Literature [4] aims to solve the sparseness of data, and establish a new recommendation model through collaborative training and combining relevant recommendation content. Literature [5] improves the recommended accuracy based on the stacked noise reduction autoencoder.

This paper proposes the LSCRS model (Location and Similar Comments Point-of-interest Recommendation System for Users). In this model, first, the user comment data is vectorized. The calculation of similarity of comments is realized, so that users' potential points of interest can be discovered. Next, this article uses the potential probability model to find the potential region with the
highest probability for each point of interest. Finally, the area to which the user belongs is obtained through the user's current location information. Use the potential interest points of the user obtained in the first step to find similar interest points in the area. This paper uses the bag-of-words model to get the similarity between different points of interest. Sort the interest points in the recommendation list according to the similarity, remove the top K recommended content, and finally get the Top-K recommendation.

2. LSCRS method
The frame of LSCRS Method is shown in Figure 1:

![Figure 1. LSCRS model framework.](image)

2.1. Regional division of space projects
In reality, users may visit multiple different points of interest in a certain geographic area. Therefore, this article divides the geographic space into R regions, and uses multiple distributions \( \theta_u \) on geographic spatial locations to simulate the distribution of users in R regions (as shown in Figure 2).

![Figure 2. Regional division model.](image)

Given hyperparameters \( \gamma, \tau \) and observable data \( v, lv \). The purpose of this article is to calculate the potential variables \( \theta, \varphi \). For each user's potential area \( r \), the posterior distribution is used instead of expressing it as a parameter to be estimated. Since this posterior distribution is difficult to calculate directly, this article uses the Marco chain Monte Carlo (MCMC) method for sampling.

Sample the region \( r \) according to the following posterior probability:

\[
p(r | r \neq u, v) \propto \frac{n_{u,v} + \gamma_r}{\Sigma_r (n_{u,r} + \gamma_r)} \frac{n_{u,v} + \tau_v}{\Sigma_v (n_{r,v} + \tau_v)}
\]  

(1)
Among it, \( n_{u,r} \) is the number of times the user \( u \) has visited the area \( r \), and \( n_{u,r} \) is the number of times the interest point \( v \) is generated from the area \( r \).

For simplicity and convenience, this article sets the hyperparameters as \( \gamma = 50/R \) and \( \tau = 0.01 \). Then in each iteration, use equations (2) and (3) to update the parameters of the Gaussian distribution. A total of about 3000 iterations are required. The parameters \( \theta \) and \( \varphi \) are estimated as follows:

\[
\theta_{u,r} = \frac{n_{u,r} + \gamma_r}{\sum_r(n_{u,r} + \gamma_r)}
\]

(2)

\[
\varphi_{r,v} = \frac{n_{r,v} + \tau_v}{\sum_v(n_{r,v} + \tau_v)}
\]

(3)

After the iteration is completed, a Gaussian distribution is applied to each region \( r \) and the position of the interest point \( v \) is represented by \( l_v \sim \mathcal{N}(\mu_r, \Sigma_r) \) as follows (3):

\[
p(l_v | \mu_r, \Sigma_r) = \frac{1}{2\pi \sqrt{||\Sigma_r||}} e^{-\frac{(l_v - \mu_r)^T \Sigma_r^{-1} (l_v - \mu_r)}{2}}
\]

(4)

Through the processing of implicit semantic model, finally, each point of interest is divided into the area with the greatest probability to which it belongs.

2.2. Analysis of review similarity

Through statistical analysis of user comment data in the recommendation system, it is found that the comment data is mainly long text. Usually short text is handled very differently from long text. Therefore, this article uses sentence vectors (Doc2vec). It includes two variants of the distributed memory model of sentence vectors (PV-DM) and the distributed bag-of-words model of sentence vectors (PV-DBOW). This article uses PV-DBOW. For a given sentence, \( w_1, w_2, ..., w_T \) represent the words in the sentence, and \( D \) represents the sentence. The objective function \( g(w) \) of the PV-DBOW model is to maximize the average logarithmic probability:

\[
g(w) = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log(w_{t+j} | w_t, D)
\]

(5)

In formula (5), \( c \) is the number of training texts. The larger the value of \( c \), the higher the accuracy of the model may be. The PV-DBOW model uses the hierarchical softmax function to define \( p(w_{t+j} | w_t, D) \).

Using PV-DBOW training sentence vector model, each comment text is represented as a \( K \)-dimensional real value vector. Using the calculation of similarity, the calculation method of similarity between comment sentences can be realized.

2.3. Similarity analysis of space items

In the data set used in this article, most of the content descriptions of points of interest are based on tags, that is, they are composed of multiple nouns and are disordered. According to the characteristics of this data, this paper adopts the Bag-of-words model (Bow model).

The formula of cosine similarity is shown in formula (6):

\[
\cos \theta = \frac{\sum_{i=1}^{n} (x_i y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \sqrt{\sum_{i=1}^{n} (y_i)^2}} = \frac{a \cdot b}{||a|| \cdot ||b||}
\]

(6)

In formula (6), \( a \) and \( b \) represent two high-dimensional vectors, and \( x_i \) and \( y_i \) represent the value of each dimension in phase \( a \) and \( b \), respectively.

3. Experimental results and analysis

In this section, first explain the analysis method of the model, the algorithm for comparison and the setting of the experiment, and then show the results of the experiment.

3.1. Dataset description

This experiment uses Yelp's real-world data set. The details of this data set are shown in Table 1.
Table 1. Yelp data description.

| Dataset                  | Yelp |
|------------------------|------|
| Number of users        | 4157 |
| Number of points of interest | 27696 |
| Number of registered records | 157657 |
| Time span              | 2009.04-2015.01 |

In the Yelp data set, 4157 users from four cities are included, as well as comments on points of interest from these users. A total of 27,696 points of interest and 1,57657 registration records are included. Each registration record includes the user's ID, point of interest ID, location of the point of interest, content information of the point of interest, comment of the point of interest, and registration time. Note that the exact recording time is not included in this data set, just an approximate date (for example: 2015-01-02).

3.2. Algorithm comparison and analysis methods of this model

This article compares LSCRS model with UPS-CF algorithm [7] and CKNN algorithm [8]. In order to fully evaluate the performance of the model, follow the method framework proposed in [6][9][10] to calculate Accuracy@k. If the interest point v appears in the result of top-k, define the value of hit@k for a single test sample as 1, otherwise it is 0.

For the overall Accuracy@k, use all test examples to define:

\[
\text{Accuracy@k} = \frac{\#\text{hit@k}}{|D_{\text{test}}|} \tag{7}
\]

Among it, \#hit@k represents the number of hit@k in all test samples, and |D_test| represents the number of all test samples.

3.3. Accuracy analysis

In order to show the improvement of the accuracy of the experiment in this article. The comparison result of LSCRS algorithm with CKNN and UPS-CF is shown in figure3:

![Figure 3](image-url)  

Figure 3. Comparison of accuracy between LSCRS algorithm and CKNN and UPS-CF recommended algorithms.

It can be seen from the above table that the recommended accuracy of the LSCRS algorithm is greatly improved. It is precisely because of the introduction of geographic location parameters that the recommendation system makes it easier for the recommendation system to find the place that the current user wants to go. Reducing the geographic scope makes the recommendation more precise. This article uses user comment data as the user's historical behavior, which increases the diversity of recommended content. It is not limited to making recommendations to users based on the content of points of interest.

For the LSCRS algorithm, this paper selects users with different numbers of comments for testing, showing the accuracy of the final recommendation results for different categories of users. As shown in Figure 4:
Figure 4. Recommendation accuracy rate of different comments.

It can be seen from figure 4 that for different TOP-K, the more historical comments of each user in the LSCRS algorithm, the higher the accuracy of recommendation. This shows that the more sufficient the user's historical record, the more basis for reference, and the more accurate the user's interest can be captured.

In order to observe the clustering of regions, this article divides regions into different numbers of categories to observe the experimental results, as shown in Figure 5:

Figure 5. Comparison of accuracy rates in different regions.

It can be observed from figure 5 that as the number of regions increases, the accuracy decreases. The best parameter setting recommended by points of interest is R = 10. This shows that R represents the complexity of the model. When R is too small, the model's ability to describe data is limited. When R exceeds a certain threshold, the model has enough expressive power to process the data. Therefore, further increasing the value of R has little effect on the model. Similarly, the increase of R will increase the time cost of training, so this article sets the value of R to 10.

4. Conclusions

Since personalized recommendation based on individuals has become a research hotspot today, this paper proposes a new LSCRS model. Firstly, Doc2vec is used to vectorize user comments. Secondly, the Bow model is used to match the similarity of spatial items, and the potential probability model is used to divide the geographical area of interest points, to find points of interest that match the user's potential interests through the user's historical comment data, determine the user's area through the user's location, and finally find the points of interest that the user may visit in this area, and then get Top-K recommended items.

References

[1] Pan Y T, He F Z, Yu H P. Social recommendation algorithm using implicit similarity in trust [J]. Chinese Journal of Computers, 2018, 41(1):65-81. (In Chinese)

[2] Kong X X, SU B C, Wang H Z. Research on the modeling and related algorithms of label-weight rating based recommendation system [J]. Chinese Journal of Computers, 2017,
40(6):1440-1452. (In Chinese)

[3] Shmueli E, Kagian A, Koren Y, et al. Care to comment: recommendations for commenting on news stories[C]//Proceedings of the 21st International Conference on World Wide Web. Lyon, France, 2012: 429-438

[4] Wang H, Wang N, Yeung D Y. Collaborative deep learning for recommender systems[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 1235-1244

[5] Wu Y, DuBios C, Zheng A X, et al. Collaborative denoising auto-encoders for top-N recommender systems[C]//Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. San Francisco, USA, 2016: 153-162

[6] Bo Hu and Martin Ester. Spatial topic modeling in online social media for location recommendation[C]. In RecSys. 2013. 25–32.

[7] Jie Bao, Yu Zheng, and Mohamed F Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data[C]. In SIGSPATIAL. 2012:199–208

[8] Gregory Ference, Mao Ye, and Wang-Chien Lee. Location Recommendation for Out-of-town Users in Location-based Social Networks[C]. In CIKM. 2013. 721–726.

[9] Weiqing Wang, Hongzhi Yin, Ling Chen, Yizhou Sun, Shazia Sadiq, and Xiaofang Zhou. Geo-SAGE: A Geographical Sparse Additive Generative Model for Spatial Item Recommendation[C]. In KDD. 2015. 1255–1264.

[10] Hongzhi Yin, Xiaofang Zhou, Yingxia Shao, Hao Wang, and Shazia Sadiq. Joint Modeling of User Check-in Behaviors for Point-of-Interest Recommendation[C]. In CIKM. 2015. 1631–1640.