Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2290–2303, DOI 10.12694/scpe.v25i4.2240

THE IMPACT OF SIMILARITY FUSION BASED TRAVEL INTEREST POINT RECOMMENDATION ALGORITHM IN YOUTH USERS' STUDY TOURS

XINLI XING AND DAOJUN WANG*

Abstract. Study tours for adolescent users are somewhat contemporary and traditional methods, such as questionnaires cannot meet their psychological expectations. To bring a better experience to teenagers' study tour, a Latent Dirichlet Allocation (LDA) theme model was used to mine teenage users and their interest points, and then the similarity between LDA and the check-in matrix was calculated and fused. Based on this, an RT-CNN model was built for deep feature extraction of review information, and point-of-interest recommendation was performed by fusing similarity, check-in behavior, and geographic location. The RT-CNN model had an accuracy of 92.7%, a recall of 87.1%, a Mean Absolute Error (MAE) value of 4.2%, a Root Mean Square Error (RMSE) of 4.8%, and F1 values of 89.2% and 88.7% in the two datasets. The new model in this experiment has high accuracy in making interest point recommendations and has a good overall performance.

Key words: Similarity fusion; Dynamic prediction; RT-CNN; Interest points; LDA topic model

1. Introduction. More traditional methods such as questionnaires are generally used in the study of study tours for young users [1-2]. For young people nowadays, these travel recommendations are no longer representative. A point-of-interest recommendation is used to analyze teenage users' preferences through data mining of their personal information, browsing history, and location information, which can help teenage users to search for new locations of interest [3-5]. The point-of-interest recommendation for teenage users can help them to choose their favorite locations among countless locations, which improves the experience of teenage users and reduces the feeling of blindness and helplessness in unfamiliar locations [6]. Study tours are organized by schools based on regional characteristics, age of students, and teaching needs of various subjects to allow students to leave the campus through group trips and centralized accommodation and meals. Their experience of collective lifestyle and social public morality can be increased in this way. Therefore, it is very important to choose an appropriate research location and content. Dynamic and intelligent recommendations are proposed based on artificial intelligence to meet the psychological expectations of young users for sports research travel, combined with the emotional preferences and geographical location of young users. It is hoped to use intelligent recommendations to help students improve their experience and expectations of study tours. The LDA topic model is used to mine teenage users and their points of interest and topics on the text information of teenage users aggregated into different documents to improve the accuracy of the point-of-interest recommendation algorithm in the study tours of teenage users. Then the similarity between the check-in matrix and the LDA topic model is calculated and fused. An RT-CNN model is built to deeply extract content such as contextual sentiment and semantics from the comment information of teenage users. Travel interest point recommendations are achieved by mining the interest preferences, emotional tendency and location information of teenage users. Then the influence of similarity, teenage users' check-in behavior, and the geographic location factors they are located are fused. It is hoped to provide a practical travel interest point recommendation method for youth study tours in Qingdao.

The existing interest point recommendation algorithms have achieved good research results in mining the evaluation information, geographic location, and other information of adolescent users. However, these technologies still suffer from some issues, such as inaccurate check-in matrix information for teenage users. LDA topic model and RT-CNN are used to mine users and their interest points to improve the accuracy of interest point recommendation algorithms in sports research travel for adolescent users. It is hoped to

*Department of Physical Education, Qingdao Agricultural University, Qingdao 266109, China (wangdaojun1973@126.com)

provide a practical tourism interest point recommendation method for youth study tours in the Qingdao area. The article consists of four parts. The first part provides a review of the research on tourism interest point recommendation algorithms. Next a detailed explanation of the methods used is provided, and corresponding models are established. The third part is the validation of the established recommendation algorithm. Finally, there is a summary of the entire article and future prospects.

2. Related works. Data sparsity and cold start are the same problems that point-of-interest recommendation algorithms need to face as traditional recommendation algorithms. In addition, point-of-interest recommendation algorithms also have difficulties in using contextual information, such as the inability to analyze the geographic location of the movement [7-9]. Meanwhile, it is a difficult problem to mine the valid information in the sparse check-in data of teenage users and making recommendations for point-of-interest recommendation algorithms. The personalized recommendation has high academic research value and commercial value, so experts at home and abroad have conducted focused and in-depth research on personalized recommendation techniques. Most recommendation methods aim to improve the recommendation accuracy. Lin Y et al. proposed to design diversity recommendation methods using scenic features. This method could classify errors and optimize themes. The accuracy and diversity of this recommendation method were confirmed in real tourism datasets [10]. Apriori can be used to establish recommendation algorithms in the tourism industry. Yang S et al. designed an intelligent recommendation system based on the Internet and artificial intelligence. This system introduced Apriori for user behavior analysis, resulting in customized recommendation content. This system had been validated after method validation, which was better than traditional recommendation methods such as decision trees and linear regression [11]. Clustering algorithms can perform clustering analysis on textual information to provide users with decision-making. S Ding et al. confirmed that Peak Density Clustering (DPC) could effectively identify cluster centers. It could divide boundaries based on the relationship between cluster centers and surrounding points after the improvement of the method. The results on different datasets indicated that the improved method could effectively control runtime and improve clustering accuracy [12]. Context-aware methods based on knowledge graphs could be applied to recommend preference items to users, and the integration of auxiliary information and collaborative filtering techniques in the recommendation algorithm was a good improvement in the performance of the algorithm [13].

The performance of recommendation algorithms can be improved to some extent by text or rating information mining, but not by deep feature mining. So topic models or deep learning techniques are proposed to use for text mining, such as CNN, LDA, and Restricted Boltzmann Machine (RBM) and other topic models or deep learning techniques for deep mining of rating information and text documents of teenage users. The embedding models based on random forests and word vectors can achieve a rational use of point-of-interest data and can facilitate the improvement of urban functions [14]. Y Luo used the LDA model for sentiment analysis. Meanwhile, fuzzy algorithms were introduced for image mining and model analysis. Small samples were analyzed and the emotions and preferences of tourists were fully considered based on this model [15]. Xu Z constructed a recommendation model using differential privacy theory and clustering methods. This model could be used for analyzing user interests and preferences. The user's preferences formed a matrix model after training. After experimental verification, the generated comprehensive model could provide recommendations for tourism interest points with high accuracy and privacy protection capabilities [16]. In social networks, interest-based recommendations are of practical application value. Song R et al. were able to use this method to analyze user data from different dimensions. A new recommendation model was constructed to comprehensively analyze data with both temporal and spatial dimensions. In real datasets, this model outperformed existing methods [17]. Geographic information is the foundation of model application in interest point recommendation models. Vinodha R et al. integrated the obtained geographic information for analyzing user behavior patterns. Meanwhile, they fixed the missing data and displayed the areas of interest to the user. Recommendation calculation based on interest points was achieved after data processing [18].

In summary, existing interest point recommendation algorithms have achieved good research results in mining evaluation information, geographic location, and other information of adolescent users [19-20]. However, these techniques still suffer from some issues, such as inaccurate check-in matrix information for teenage users. In addition, these algorithms have not screened the comments of adolescent users, losing certain features of their comment information and a decrease in recommendation accuracy. The LDA topic model is used to

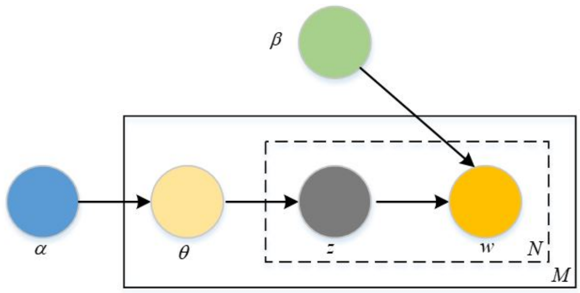


Fig. 3.1: Generation process of LDA theme model

mine users and their interest points to improve the accuracy of interest point recommendation algorithm in sports research travel for adolescent users. Meanwhile, CNN is introduced into the model for value screening of user comment information, which can accurately reflect the characteristics of interest points. The proposed method aggregates all comment information related to the point of interest and distinguishes user comment information to a certain extent. It extracts some user comments during the information extraction to reflect the characteristics of the points of interest. The algorithm proposed in this experiment can effectively address the inaccurate feature extraction caused by the lack of importance differentiation for documents. It effectively addresses some of the shortcomings of the current approach and is superior.

3. The similarity fusion and CNN-based travel interest point recommendation algorithm in youth sports research trips.

3.1. The travel interest point recommendation algorithm based on similarity fusion. In the interest recommendation algorithm, text information related to interest points and teenage users is input into the LDA topic model for the point of interest and teenage users. Their latent topic topic features can be mined and represented by a topic vector. The method used in the LDA topic model is the bag-of-words method. A document (text) is composed of words (w) that have no sequential relationship, and these words are attributed to a topic (topic) in the document with a certain probability, while the topic in the document also has a certain probability to select a word. If a document is to be generated, then the probability of occurrence of words in it is given in equation (3.1).

$$p(w \mid text) = \sum_{topic} p(w \mid topic) \times p(topic \mid text) \tag{3.1}$$

In the same document that can contain more than one topic, the topics are output in a probability matrix of the LDA model. The topic distribution of words never can be obtained, and then the similarity of the document is calculated based on the topics. The model generation process is shown in Fig 3.1. First a random topic distribution θ is selected, then a random topic from z is selected, in the word probability based on the random topic z to generate words w. Word w is generated based on the probability of words in the random topic z.

In Fig. 3.1, α and β are corpus-level parameters, and θ denotes a topic vector of dimension$K$. Both the topic z and $word$ are word variables. The topic z is generated by the topic vectorθ. The word w is generated jointly by the topic z and the parameter β. Then the joint probability in equation (3.2) can be obtained.

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_\theta \mid z_\theta, \beta) \tag{3.2}$$

Combining Fig. 3.1 with equation (3.2) yields the probability generation for the LDA model in Fig 3.2.

In the LDA model, α and β need to be trained to obtain the parameters. α is used to represent the probability distribution of "document-topic". β is used to represent the probability distribution of "topic-word".

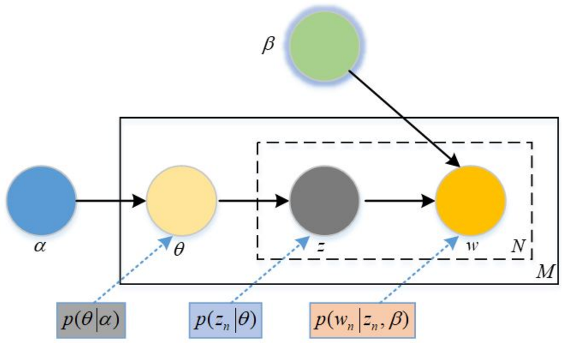


Fig. 3.2: Probability generation process of LDA model

For the training of the parameters α and β, the learning algorithm usually chosen is Gibbs sampling method, which first requires randomly setting the topics to which the words in the document belong. Then the probability of each word in the document belonging to the topic is updated according to equation (3.2), while satisfying the randomness of the word selection. Finally, the operations in equation (3.3) are continuously executed until the iteration condition is satisfied.

$$p(z_{i'} = k \left| z_{-i'}, w \right.) \propto \frac{n_{m,-i'}^{(k)} + \alpha_k}{n_{m,i'}^{(\cdot)} + V\alpha_k} \times \frac{n_{m,-i'}^{(s)} + \theta_s}{\sum_{s=1}^{V}(n_{k,-i'}^{s} + \beta_s)} \tag{3.3}$$

In equation (3.3), z_{-i}¡ is the distribution of topics excluding subscript i. The word distribution under the topic can be obtained after the iterations in equation (3.3) based on equation (3.4).

$$\theta_{k,s} = \frac{n_k^{(s)} + \theta_s}{\sum_{s=1}^{V}(n_k^{(s)} + \beta_s)} \tag{3.4}$$

The topic distribution of the document can be obtained after the iterations in equation (3.3) based on equation (3.5).

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{n_m^{(\cdot)} + V\alpha_k} \tag{3.5}$$

In the LDA-based phase-metric model, it is first necessary to aggregate the teenage users' comments and description information related to the same interest point into a single document. Next, all the labeled and commented information related to the interest point that has been checked in by the same teenage user is pooled into one document. Then a large collection of documents can be obtained, where one document corresponds to one teenage user or one point of interest. In this experiment, the number of interest points given and the topics implied in teenage users is K. Then the topic distribution of an interest point θ_p and the topic distribution of teenage users θ_u should be computed. θ_p and θ_u denote vectors of dimension K. Each dimension of the vector represents the probability size generated by the teenage user or interest point under the corresponding topic. The LDA model is used to train the topic features of each teenage user or point of interest. Equation (3.6) shows the process of solving the similarity of the teenage user's nearest neighboring topic features.

$$sim(\theta_{u_i}, \theta_{u_j}) = \frac{\sum_{m=1}^{K}(w_m^{(u_i)} \times w_m^{(u_j)})}{\sqrt{\sum_{m=1}^{K}((w_m^{(u_i)})^2 \times \sum_{m=1}^{K}((w_m^{(u_j)})^2)}} \tag{3.6}$$

In equation (3.6), θ_{u_i} is the topic vector of teenage user i. θ_{u_j} is the topic vector of interest point j. The probability that teenage user i and interest point j belong to topic m is denoted as $w_m^{(u_i)}$ and $w_m^{(u_j)}$, respectively. The similarity measure based on the teenage user check-in matrix and the LDA topic model is divided into the following steps. First, the information documents related to teenage users and points of interest are input. The teenage user documents or points of interest documents that need to be processed in the document set are word-sorted and the words are converted into word frequency vectors. Next, the relevant parameters of LDA are set, the processed documents are put into LDA for training, and the model output is transformed. For any teenage users u_i and u_j, the similarity is calculated using equation (3.6). Finally, the similarity matrix is output. It is supposed the set of teenage users recommended in the points of interest is denoted as $U_{POI} = \{u_1, u_2, ..., u_m\}$, the set of points of interest is denoted as $V_{POI} = \{u_1, u_2, ..., u_n\}$, a check-in matrix of size $N \times M$ is denoted as R, and the check-in term is $r_{i,j}$. In equation (3.7), the cosine similarity formula is used to carry out the similarity measure of the teenage users obtained based on the check-in matrix.

$$sim(x_a, x_b) = \frac{\sum_{i \in N(a) \cap N(b)} n_{a,i} \cdot n_{b,i}}{\sqrt{\sum_{i \in N(a)} (n_{a,i})^2 \cdot \sum_{i \in N(b)} (n_{b,i})^2}} \tag{3.7}$$

In equation (3.7), $n_{a,i}$ is the check-in frequency of teenage user a to interest point i and $n_{b,i}$ is the check-in frequency of teenage user b to interest point i. The set of nearest neighbors of teenage users can be obtained at $U_l = \{u_{l_1}, u_{l_2}, ..., u_{l_m}\}$ after the similarity is calculated using the LDA model. The set of nearest neighbors of teenage users based on the check-in data similarity measure is $U_t = \{u_{t_1}, u_{t_2}, ..., u_{t_n}\}$. The final set of nearest neighbors is $U_b = U_t \cup U_l$. The set of nearest neighbors that can be derived by both methods is denoted as $U_j = U_t \cap U_l$ and its weight value will be increased when performing the calculation of predicted visit probability. In equation (3.8), the similarity of teenage users can be set.

$$sim(u_a, u_b) = \mu sim_l(u_a, u_b) + (1 - \mu) sim_t(u_a, u_b) \tag{3.8}$$

In equation (3.8),$sim_l(u_a, u_b)$ is the similarity calculated using the LDA model, $sim_t(u_a, u_b)$ is the similarity calculated based on the check-in data similarity measure, and μ is the parameter of size 0.6. Dynamic prediction is introduced in this experiment to avoid the emergence of new errors, reduce the impact caused by data sparsity, and dynamically fill in the missing information for the current check-in of teenage users. The solution of the current interest point nearest neighbor set is performed using equation (3.6). The probability of access to the current missing interest point check-in information is predicted based on this nearest neighbor set. Then the probability of access is predicted. It is supposed that the current target teenage user is u, its interest point is p, M is denoted as the nearest neighbor set of the teenage user, and h denotes any teenage user of the nearest neighbor set M. Then the access prediction of h to the current interest point is denoted as $\hat{c}_{h,p}$ in equation (3.9).

$$\hat{c}_{h,p} = \begin{cases} c_{h,p} & c_{h,p} \neq 0 \\ \frac{\sum_{l \in N} sim(p,l) \cdot c_{h,l}}{\sum_{l \in N} sim(p,l)} & c_{h,p} = 0 \end{cases} \tag{3.9}$$

In equation (3.9), N is the set of h nearest neighbor interest points. $sim(p, l)$ is the similarity of interest point p to l. Then the access probability of the teenage user u to the current interest point in equation (3.10) can be obtained.

$$\hat{c}_{h,p} = \frac{\sum_{h \in v} w_{u,h} \cdot c_{h,p}}{\sum_{h \in v} w_{u,h}} \tag{3.10}$$

The v in equation (3.10) is the nearest neighbor set of teenage users, and $w_{u,h}$ is the similarity weight of u and h. The visit probability of teenage users can be predicted by using equation (3.10), and the top N interest points with the highest visit probability are selected according to the calculated visit probability size, thus generating a recommendation list of top-N for the target teenage users.

3.2. The travel interest point recommendation algorithm based on similarity fusion and CNN. In the previous study, the similarity fusion-based interest point recommendation algorithm has good recommendation interpretation, but it needs to continue to improve in terms of accuracy. CNN has better performance advantages in natural language and text processing and can be used for text or image feature extraction. Therefore, it is chosen for the optimization of the above interest point recommendation algorithm in this experiment. In general, teenage users tend to prefer interest points that are closer to their geographical location. This model can be optimized by equation (3.11). to predict the check-in interest of teenage users u_i to the geographic location l_j where no check-in is performed.

$$\min \frac{1}{2}(H \odot (R - UL^T))^2 \tag{3.11}$$

H is the check-in weight matrix, $H \in R^{M \times N}$, whose value of 1 indicates check-in and whose value of 0 indicates no check-in in equation (3.11). The regularization terms of the parameters U and L are added to the value equation (3.11) to obtain equation (3.12) to prevent the occurrence of overfitting.

$$\min \frac{1}{2} \left\|H \odot (R - UL^T)\right\|_F^2 + \frac{\lambda_u}{2} \left\|U\right\|_F^2 + \frac{\lambda_l}{2} \left\|L\right\|_F^2 \tag{3.12}$$

λ_u and λ_l are the regularization terms of parameters U and L respectively. $\|\cdot\|_F^2$ is the Frobenius parametrization, which is optimized using the gradient descent method to obtain the local optimal solution in equation (3.12). There is a correlation rule between things. If a teenage user is interested in the neighboring locations around a geographic location, then the teenage user is more likely to be interested in that geographic location. Therefore, the neighboring location weighting method can be used to fill in the corresponding vacant geographic locations in the matrix decomposition model, and equation (3.13) is the minimization of the objective function.

$$\min_{U,L} \frac{1}{2}(H \odot (R - UBL^T)^2) \tag{3.13}$$

In equation (3.13), $B = \gamma UL^T + (1 - \gamma)A^T, A \in R^{n \times n}$. γ denotes the parameter of the neighborhood location weight. The sparse data can be mitigated by fully mining the review information of teenage users. CNN can extract the latent features from the document information in the previous paper and incorporate them in the point-of-interest recommendation. In CNN, the input data is a word vector and the output data is sentiment tendency, the model contains 4 main layers, as shown in Fig. 3.3.

First of all, the first layer is the embedding layer, which can merge words from all evaluation information of a particular teenage user to form a single document. Then the corresponding word vector mapping is performed by the word vector model in the order of word occurrence to generate a word vector matrix with constant word order, as shown in equation (3.14).

$$M_i = (w_0) : (w_1) : ...(w_p) : ... : (w_{n-1}) \tag{3.14}$$

w_p is the word vector of word p in equation (3.14). The second layer is the convolution layer, which can be used to extract new features from the input M_i by convolution operations. The new features generated by each convolution are shown in equation (3.15).

$$T_y = f(M_i \otimes F_q + b_q) \tag{3.15}$$

f is the activation function, $\otimes$ is the convolution operation, F_q denotes the filter, and b_q is the bias term of F_q in equation (3.15). The third layer is the pooling layer, which can be used to generate new features by extracting the maximum contextual feature vector via max pooling operation. This operation can effectively handle evaluation texts of different lengths and can compress features to reduce their size and ensure that only the main features are extracted. It can downgrade the complexity of the computation and avoid overfitting of the model. The representation of pooled features is shown in equation (3.16).

$$D_j = T_y(Max(allfeature) \tag{3.16}$$

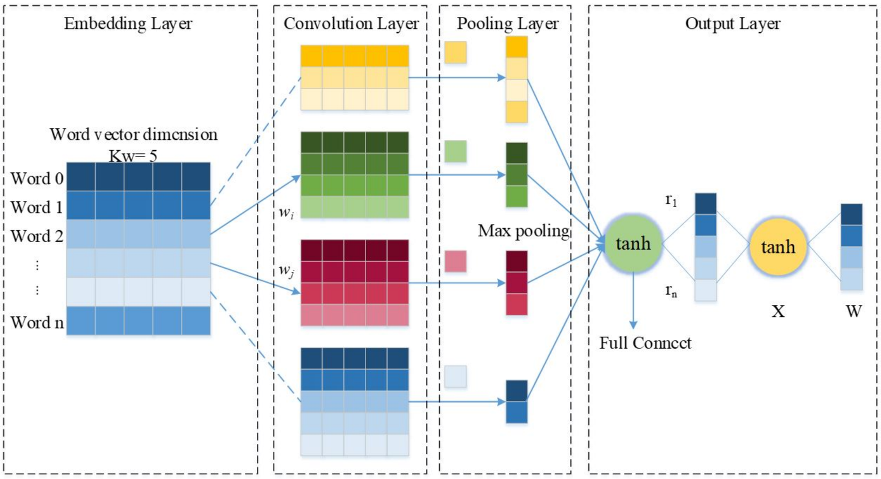


Fig. 3.3: Structure of CNN

The fourth layer is the output layer, which can input the sentiment feature vector extracted in the third layer to the Softmax function for the calculation of sentiment prediction probability. It is compared with the standard experimental data to get the error. The error is passed through the method of back propagation and gradient descent to update the parameters. In equation (3.17), CNN can deeply extract the latent features of the location interest point evaluation text and the probability of the teenage user posting the evaluation is defined using the Softmax function.

$$P(\varphi_{il} = 1 \,|u_i\,, c_l) = \frac{e^{(u_i^T \cdot C \cdot CNN(W,C_l))}}{\sum_{C_q \in C} e^{(u_i^T \cdot C \cdot CNN(W,C_q))}} \tag{3.17}$$

In equation (3.17), φ_{il} is whether the teenage user has published a rating, c_l denotes the set of documents evaluated, C means the interaction matrix used to analyze whether the teenage user has published a rating, $CNN(W, C_l)$ is the features of the evaluation text extracted by the CNN, and W refers to the internal weight of the CNN. There is a correlation between the output values of the Softmax logistic regression function, and the sum of their probabilities is equal to 1. It is necessary to transform equation (3.17) into the objective function in equation (3.18) to obtain the teenage user's potential feature vector.

$$\sum_{i=1}^{n} \sum_{c_q \in C_{ui}} \log P(\varphi_{iq} = 1 \,|u_i\,, c_q) \tag{3.18}$$

Similarly, the probability function for the correlation between geographic location and evaluation is given in equation (3.19).

$$P(\varphi_{jk} = 1 \,|l_j\,, c_k) = \frac{e^{(l_j^T \cdot P \cdot CNN(W,C_k))}}{\sum_{C_q \in C} e^{(l_j^T \cdot P \cdot CNN(W,C_q))}} \tag{3.19}$$

If the potential eigenvectors of locations are to be obtained, it is necessary to transform equation (3.19) into the objective function in equation (3.20).

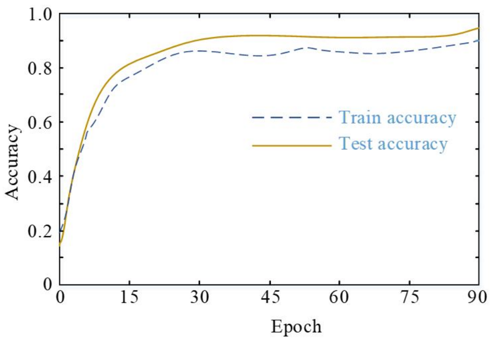
$$\sum_{j=1}^{m} \sum_{c_q \in C_{lj}} \log P(\varphi_{jq} = 1 \,|l_j\,, c_q) \tag{3.20}$$

The RT-CNN model incorporates the similarity SIM, geolocation influence B, location latent features L, check-in behavior R, teenage user sentiment tendency S, and teenage user latent features U. The location latent features L, teenage user sentiment tendency S, and teenage user latent features U are learned from CNN. The implementation steps of the RT-CNN model include the following process. First, $R, C_{u_i}, u_i \in U, C_{l_j}, l_k \in L$ are input in the model. Then SIM, W, U, L, P, C are randomly initialized. Next, CNN is used to obtain the S value, and the geographic location influence B is calculated by U, L. It is judged whether the function reaches the convergence state, otherwise parameters are updated and adjusted by back propagation method until the requirement is satisfied. Finally, top-N interest points are generated. Because this experiment is aimed at the youth sports research tourism in the Qingdao area, the potential features of geographic location and location in the model are chosen to analyze the tourist attractions in the Qingdao area.

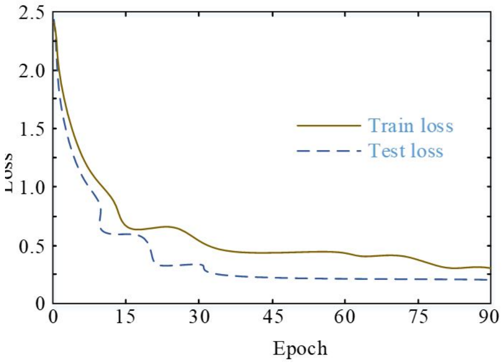
4. Simulation analysis of travel interest point recommendation model. The similarity fusion-based travel interest point recommendation algorithm was trained in the LDA model into the parameters. The dataset chosen was Yelp to evaluate the effect of the training of the method. The dataset provided by the 6th Yelp Challenge launched by the Yelp website in 2015 was used. The dataset file used in this article contains three JSON format documents, which are about users, points of interest, and comment information. The information contained in the document includes the user's ID, rating, comment text, category of points of interest, latitude and longitude, etc. A preprocessing method similar to ConvMF was adopted for the original dataset. Users who had accessed information less than 5 times and interest points that had been visited less than 5 times were removed in this article due to the extreme sparsity of the Yelp dataset. The maximum length of a text document with user and interest points set should not exceed 300. All stop words were deleted. TF-IDF was used to calculate the weight of each word in the document, and words with weights not exceeding 0.5 were deleted. The top 8000 words with weight values were selected as the final vocabulary. The ratio of training set to test set was 7:3. The model was put into the training set and the test set for training, respectively. The training results of the travel interest point recommendation algorithm based on similarity fusion are shown in Fig. 4.1. The parameters of the comparison algorithm were set to achieve the best algorithm recommendation effect to demonstrate the effectiveness of the algorithm in this article. In this experiment, the dimensionality of the potential user and interest point matrices in matrix factorization was set to 50, and random initialization was performed. The regularization parameters α_U and α_V of the recommendation algorithm were set to 0.002 and 100, respectively. The word vector dimension kw of the LDA topic model was set to 200. The user and interest point text documents were set to a maximum length of 300 in this experiment to better train the CNN model. The dimension of word embedding was set to 200 in the experiment. The windows in the convolutional layer were set to 3, 4, and 5, respectively, to better extract document features. The number of windows of each width was 100. Dropout was used for mapping in the experiment to prevent overfitting, with a loss rate of 0.3. Accuracy is the proportion of all correctly predicted (including positive and negative classes) to the total, used to evaluate the generalization ability of a model, that is, the performance of the model. The higher the accuracy, the stronger the generalization ability of the model, indicating better performance. The accuracy of the similarity fusion-based recommendation algorithm in the test set was 93.7%, which was higher than the value in the training set. The loss function value of the method was 0.48 in the test set, which was lower than the value in the training set. The method could get better parameter optimization in the test set after learning from the training set.

The RT-CNN model was also trained on both training and test sets in Fig. 4.2 to validate the similarity fusion and CNN-based travel interest point recommendation algorithm. The accuracy of the CNN-based recommendation algorithm in the test set was 94.8%, which was higher than the value in the training set. The loss function value of the method was 0.13 in the test set, which was lower than the value in the training set. The method could get better parameter optimization in the test set after learning from the training set.

The accuracy of the RT-CNN model might be affected by the word vector dimension and the number of convolutional kernels. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were used to verify the influence of word vector dimension and the number of convolutional kernels to improve the application of the mode in Fig. 4.3. MAE is the average absolute value of the deviation between all individual observations and the arithmetic mean, which can better reflect the actual situation of prediction error. The smaller the MSE, the better performance of the model. RMSE represents the sample standard deviation of the difference

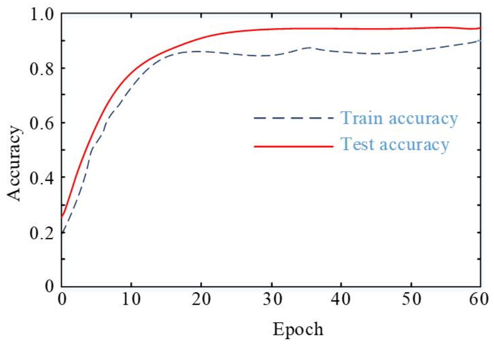


(a) Accuracy

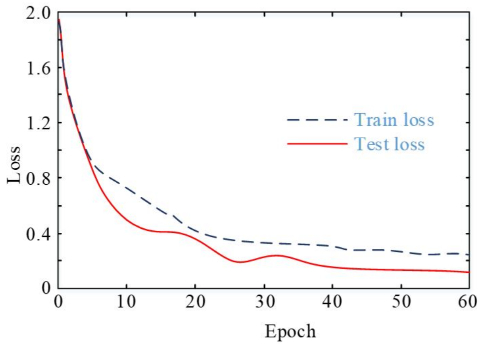


(b) Loss Value

Fig. 4.1: Accuracy and loss results



(a) Accuracy

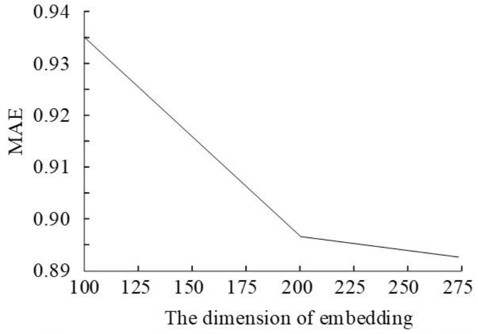


(b) Loss Value

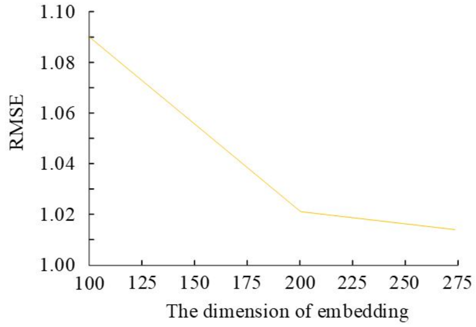
Fig. 4.2: Accuracy and loss results of RT-CNN

between predicted and observed values. RMSE can indicate the dispersion of the sample. When dealing with non-linear fitting, the smaller the RMSE, the better performance of the model. When the size of the word vector dimension was 200, the model's MAE and RMSE values were 0.897 and 1.021, while the model synthesis effect and complexity were the best. When the number of convolutional kernels was greater than 250, the MAE and RMSE values of the model gradually increased. At this point, it might be due to too much feature selection, which led to the occurrence of an overfitting phenomenon, resulting in a decrease in the comprehensive effect of the model. Therefore, the size of the word vector dimension was set to 200 and the number of convolutional kernels was set to 250 in the subsequent experiments.

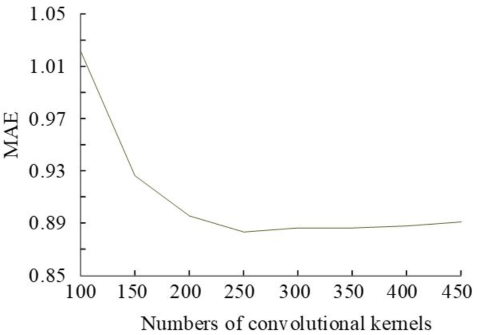
Regarding the performance evaluation of the above point-of-interest recommendation algorithm, the more widely used evaluation metrics were selected, which included the Precision (P), Recall (R), and the ratio of the two (F1). P is the proportion of correctly predicted positive classes to all predicted positive classes. A high P means that it is definitely positive as long as it is recognized as positive. It focuses on dividing indistinguishable samples into negative samples, examining whether the identified positive samples are reliable. R is the proportion of correctly predicted positive classes to all actually positive classes. A high R means that it can be recognized as long as it is positive. It focuses on dividing indistinguishable samples into positive samples, examining whether they are sensitive to positive samples. The F1 value is a measure of the classification problem



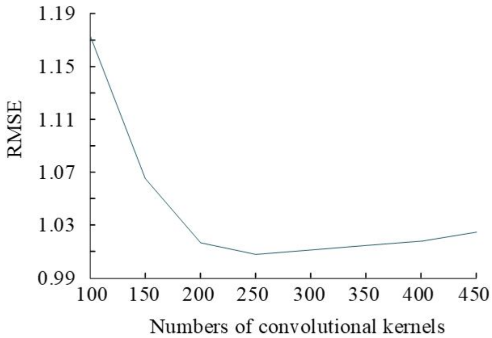
(a) MAE of different dimenssion of word embedding



(b) RMSE of different dimenssion of word embedding



(c) MAE of different number of convolutional kernels

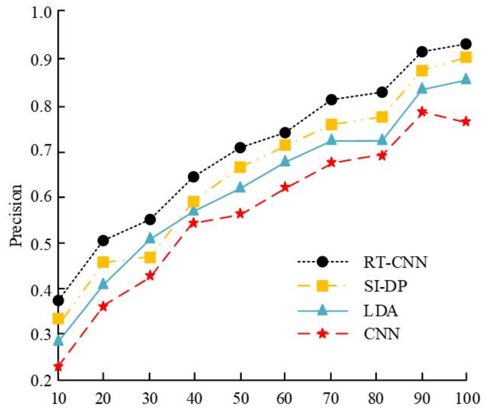


(d) RMSE of different number of convolutional kernels

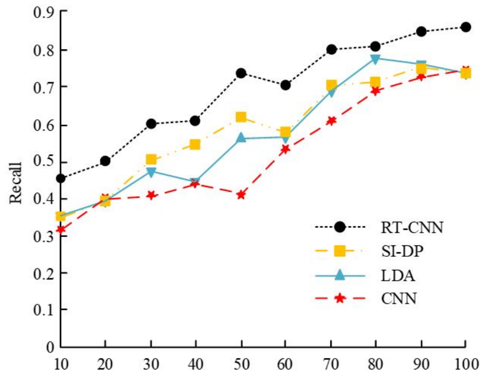
Fig. 4.3: Accuracy and loss results of RT-CNN

and can be used to evaluate the quality of a model. It is the harmonic mean of P and R, with a maximum of 1 and a minimum of 0. The higher the F1 value, the better the performance of the model. The difference in the metrics between the RT-CNN model, literature model [21], LDA model, and traditional CNN model were compared in Fig. 4.4. The R of the RT-CNN model was 92.7%, which was higher than the other models. The R of this model was 87.1%, which was the highest among all models, indicating the superior performance of the RT-CNN model. The MAE and RMSE metrics could be used to validate the P and R metrics with confidence in Fig. 4.5. The average MAE value of the RT-CNN model was 4.2% and its average RMSE value was 4.8%, which were lower than the literature model, LDA, and the traditional CNN, proving the results were accurate. The reason was that the LDA topic model was utilized to mine users and their interests. Then the similarity between LDA and check-in matrix was calculated and fused. On this basis, an RT-CNN model was established to extract deep features from comment information, and interest point recommendations were made by integrating factors such as similarity, check-in behavior, and geographic location. This effectively improved the accuracy of the model and was superior to existing methods.

To further verify the effect of RT-CNN model P and R, some data in Yelp dataset were selected as datasets 1 and 2 to verify whether the F1 value of the model was consistent with the results of P and R in Fig. 4.6. The F1 values of the RT-CNN model were 89.2% and 88.7%, which were higher than other algorithms and consistent with the above results of P and R metrics, which further verified the accuracy of the experimental results. The reason was that the CNN was introduced into the model for value filtering of user comment information, which could accurately reflect the characteristics of interest points. The proposed method aggregated all comment

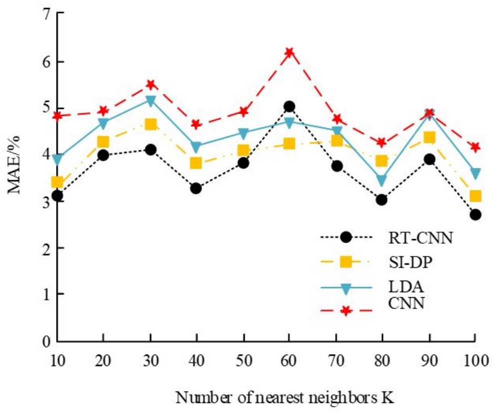


(a) Precision

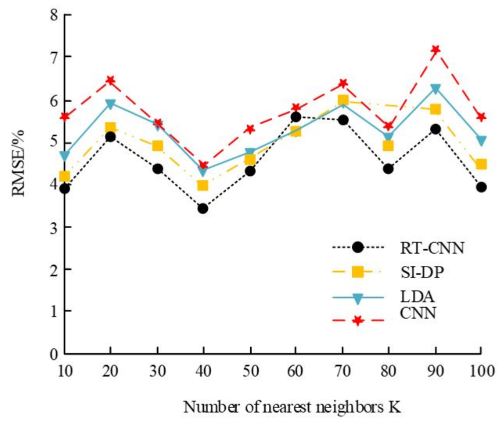


(b) Recall

Fig. 4.4: Precision and recall results of RT-CNN



(a) MAE

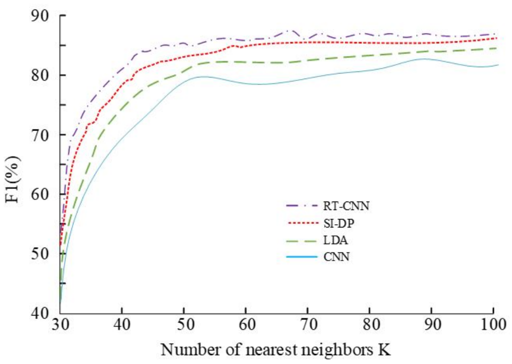


(b) RMSE

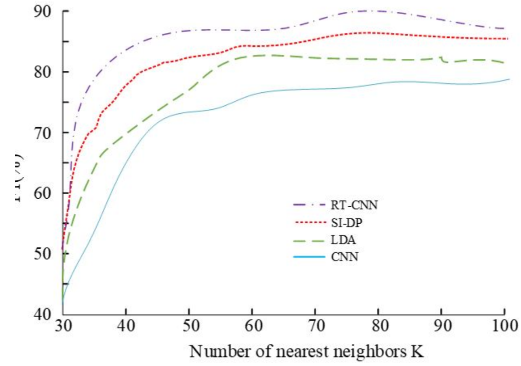
Fig. 4.5: Comparison of MAE and RMSE

information related to interest points and distinguished user comment information to a certain extent. This was beneficial for improving the recommendation accuracy of the model and fully reflecting limited data information.

Precision-Recall (PR) curve is a comprehensive judgment of precision and recall. This study compared the performance of the RT-CNN model, literature model, LDA, and traditional CNN in Fig. 4.7. PR comprehensively considers the situations of correct classification and incorrect classification. It has high practicality in extremely imbalanced data. But when the proportion of positive and negative samples is different, it will show significant differences. Therefore, PR can serve as a reference for model performance evaluation. The area under the PR curve of the RT-CNN model were higher than that of literature model, LDA, and traditional CNN, which proved that the travel interest point recommendation algorithm in this experiment had better accuracy. The reason was that the algorithm proposed in the experiment extracted some user comments during the information extraction, which were used to reflect the characteristics of the points of interest. This algorithm could effectively solve the inaccurate feature extraction caused by the lack of importance differentiation for



(a) Dataset1



(b) Dataset2

Fig. 4.6: F1 value of RT-CNN

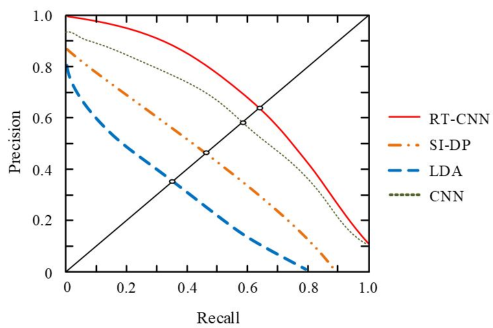
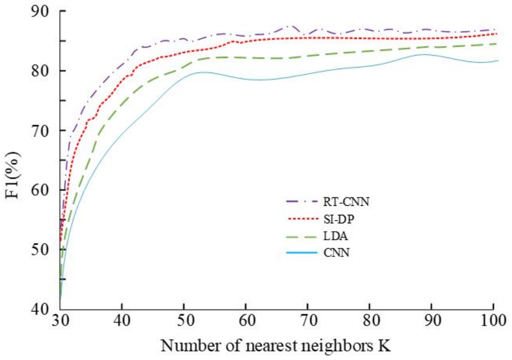


Fig. 4.7: Comparison of PR curves of different algorithms

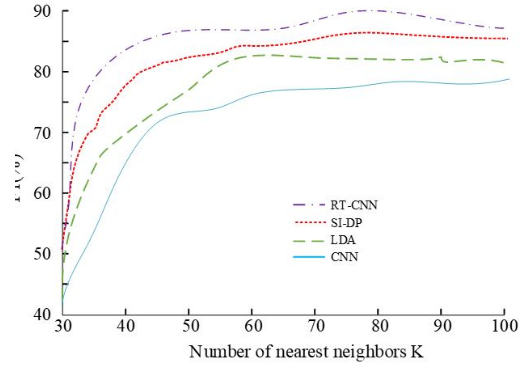
documents. Meanwhile, it only filled in nearest neighbor access when the user's nearest neighbor had missing access to a certain point-of-interest, effectively improving the execution efficiency of the algorithm.

At the same time, the Receiver Operator Characteristic (ROC) curves of the RT-CNN model, literature model, LDA, and traditional CNN were compared in Fig. 4.8. ROC is an integrated metric that reflects the sensitivity and specificity of continuous variables, and each point on the ROC curve reflects the sensitivity to the same signal stimulus. The steeper the ROC curve, the better the performance of the model. The larger the area between the curve and the horizontal axis, the better the performance of the classifier. The ROC curve can be used to determine which of the two classification methods is better, or to evaluate different parameters of the same classification method and select the best parameters. Fig. 4.8a shows the validation ROC curve plot and Fig. 4.8b shows the recognition ROC curve plot. The area under both the validation and recognition ROC curves of the RT-CNN model was higher than that of the literature model, LDA, and the traditional CNN, indicating that the RT-CNN model proposed in this experiment achieved better results.

In summary, the proposed model achieved good results in all indicators. This article studied the extraction of word features from text information in algorithms to improve the accuracy of latent representation of interest points. However, the emotional factors contained in the textual information were not explored to better fit user preferences. In addition, the features extracted by the CNN model could not be artificially interpreted, so the interpretability of the recommendation results was poor. Therefore, the performance of the model still needed further improvement. The user-interest point check-in matrix and text information to model user preferences



(a) Verification of ROC Curve



(b) Detection and Identification of ROC Curve

Fig. 4.8: ROC curve

were utilized and the interest point feature information was extracted. There was a possibility of overfitting in the proposed model. The reason was that the criteria used to select a model were different from those used to determine its applicability. Therefore, the model could be selected by maximizing its performance on certain training datasets. In future work, it needs to consider incorporating factors such as geographic location, time, popularity, and user friendly relationships from check-in data into the algorithm, and combining them with matrix decomposition to alleviate data sparsity issues.

5. Conclusion. In this study, a travel interest point recommendation algorithm was built based on similarity fusion, and this recommendation algorithm was optimized using CNN to improve the performance of the model such as accuracy. Finally, an RT-CNN travel interest point recommendation model was generated. In the experimental results, the size of the word vector dimension of the model was set to 200 and the number of convolutional kernels was set to 250. The P of the RT-CNN model was 92.7%, which was higher than other models. The R value of this model was 87.1%, which was the highest among all models. The average MAE value was 4.2% and the average RMSE value was 4.8%, both lower than the literature model, LDA, and the traditional CNN. The F1 values in both datasets were 89.2% and 88.7%. The PR and ROC curves of the model were also plotted both agree with the results in P and R. The similarity fusion-based travel interest point recommendation algorithm proposed in this experiment had better performance and better overall results after combining with CNN. Although the recommendation model in the experiment has achieved good application results, there are still some shortcomings. For example, the feasibility, innovation, and spatial complexity of recommendation algorithms are considered in reality. But the time complexity in recommendation algorithms is less considered. In the next research work, more consideration will be given to the time complexity of the algorithm. This paper will explore the emotional factors contained in textual information to better fit user preferences and further improve algorithm performance. In future work, it is necessary to consider incorporating factors such as geographical location, time, popularity, and friend relationships among users into the algorithm based on check-in data. Meanwhile, it is necessary to combine it with matrix factorization to alleviate the problem of data sparsity. It is hoped that further optimization of the algorithm can be effectively used in travel interest point recommendations for different populations.

Fundings. The research is supported by the 2021 Qingdao Social Science Planning and Research Project ”Research on the Development and Practice of Qingdao Sports Cultural Resources Based on Research and Study Travel” (No. QDSKL2101232); Shandong Province Art Science Key Project in 2020 ”Innovative path of Deep integration of Sports and cultural tourism industry in Shandong Province in the post-epidemic era” (No. YQ202008017).

REFERENCES

[1] Si, Q., Cheng, X., Du, S., Liu, J., Dong, X., Yi, F. & Xu., Z. Influence factors and promoting methods of teenagers' outdoor environment sports participation from the perspective of ecosystem theory. *Journal Of Environmental Protection And Ecology*. **22**, 829 (2021)

[2] Gao, M., Zhao, Y., Shen, Y., Yu, X., Gou, S. & Bao, Q. Which factors are most relevant to drivers' overtaking choices at two-lane highways: A comparative analysis between questionnaire surveys and driving simulation. *Transportation Research Part F: Traffic Psychology And Behaviour*. **95** pp. 202-214 (2023,5)

[3] Vinodha, R. & Parvathi, R. Content and Location Based Point-of-Interest Recommendation System Using HITS Algorithm. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **31Suppl. 1), 31** pp. 31-45 (2023,5)

[4] Liu, Y. & Wei, N. Research on location and similar comments in point - of - interest recommendation system for users. *Journal Of Physics: Conference Series*. **1774** (2021)

[5] ChenYile, L., CongGao, L., BaoZhifeng, L. & GuWanli, Z. Points - of - interest relationship inference with spatial - enriched graph neural networks. *Proceedings Of The VLDB Endowment 15(3*. pp. 504-512 (2021)

[6] Song, Y. & Dong, Y. Research on intelligent recommendation of ecotourism path based on popularity of interest points. *International Journal Of Information And Communication Technology*. **20**, 423 (2022)

[7] Wang, Y., Zhu, Y., Zhang, Z., Liu, H. & Guo, P. Design of Hybrid Recommendation Algorithm in Online Shopping System. *Journal Of New Media*. **3**, 119 (2021)

[8] Liu, Z., Ma, Y., Zheng, H., Liu, D. & Liu, J. Human resource recommendation algorithm based on improved frequent itemset mining. *Future Generation Computer Systems*. **126**, 284 (2021)

[9] Gan, M. & Tan, C. Mining multiple sequential patterns through multi-graph representation for next point-of-interest recommendation. *World Wide Web*. **26**, 1345-1370 (2023)

[10] Lin, Y., Huang, C., Yao, W. & Shao, Y. Personalised attraction recommendation for enhancing topic diversity and accuracy. *Journal Of Information Science*. **49**, 302-318 (2023)

[11] Yang, S. & Yingwei, H. Toward an intelligent tourism recommendation system based on artificial intelligence and IoT using Apriori algorithm. *Soft Computing: A Fusion Of Foundations, Methodologies And Applications*. **27**, 19159-19177 (2023)

[12] Ding, S., Du, W., Li, C., Xu, X., Wang, L. & Ding, L. Density peaks clustering algorithm based on improved similarity and allocation strategy. *International Journal Of Machine Learning And Cybernetics*. **14**, 1527-1542 (2023)

[13] Wu, C., Liu, S., Zeng, Z., Chen, M., Alhudhaif, A., Tang, X., Alenezi, F., Alnaim, N. & Peng, X. Knowledge graph - based multi - context - aware recommendation algorithm. *Information Sciences*. **595**, 179 (2022)

[14] Sun, Z., Jiao, H., Wu, H., Peng, Z. & Liu, L. Block2vec: An Approach for Identifying Urban Functional Regions by Integrating Sentence Embedding Model and Points of Interest. *International Journal Of Geo - Information*. **10**, 339 (2021)

[15] Luo, Y., Tong, T., Zhang, X., Yang, Z. & Li., L. Exploring destination image through online reviews: an augmented mining model using latent Dirichlet allocation combined with probabilistic hesitant fuzzy algorithm. *Kybernetes: The International Journal Of Systems & Cybernetics*. **52**, 874-897 (2023)

[16] Xu, Z., Hu, Z., Zheng, X., Zhang, H. & Luo, Y. A matrix factorization recommendation model for tourism points of interest based on interest shift and differential privacy. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology*. **44**, 713-727 (2023)

[17] Song, R., Li, T., Dong, X. & Ding, Z. An effective points of interest recommendation approach based on embedded meta-path of spatiotemporal data. *Expert Systems: The International Journal Of Knowledge Engineering*. **40**, 1-19 (2023)

[18] Vinodha, R. & Parvathi, R. Content and Location Based Point-of-Interest Recommendation System Using HITS Algorithm. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **31Suppl. 1** pp. 31-45 (2023,5)

[19] Naserian, E., Wang, X., Dahal, K., Calero, J. & H. Gao, A. Partition - Based Partial Personalized Model for Points of Interest Recommendations. *IEEE Transactions On Computational Social Systems*. **8**, 1223 (2021)

[20] Cai, L., Wen, W., Wu, B. & Yang, X. A coarse - to - fine user preferences prediction method for point - of - interest recommendation. *Neurocomputing*. **422**, 1 (2021)

[21] Kang, H., Ma, Y. & Si, X. An Enhanced Algorithm for Dynamic Data Release Based on Differential Privacy. *Procedia Computer Science*. **174** pp. 15-21 (2020)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Apr 13, 2023
Accepted: Apr 9, 2024