CrowdTravel: Leveraging Cross-Modal CrowdSourced Data for Fine-grained and Context-based Travel Route Recommendation

Jing Zhang[1], Bin Guo[1], Zhimin Li[1], Yan Liu[1], Zhiwen Yu[1], Qi Han[2]
1. School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, P.R.China
2. Department of Computing, Colorado School of Mines, Colorado, USA
guob@nwpu.edu.cn

***Abstract*—Travel route planning is generally a very time-consuming task due to massive travel information and diverse travel needs. In this paper, we propose CrowdTravel to automatically extract context-related scenic route recommendation through cross-modal mining and crowd intelligence extraction. First, we leverage the hybrid CNN-RNN model to learn the relationship between the photos and descriptive texts in travelogues. Second, we propose the CrowdRank algorithm to select diverse and representative photos for each scenic spot. Finally, according to users requests and particular contexts, we leverage sequential pattern mining and context filtering to generate visual- and context-based scenic routes. We conduct experiments over a dataset of 11,542 travelogues and 11,228 travel albums of eight popular scenic spots in China. Extensive experiments show the effectiveness of the proposed framework.**

***Index Terms*—Travel route recommendation, Cross-modal mining, CNN-RNN model, Crowd intelligence**

I. Introduction

With the rapid development of economy, travel is not a luxury for the minority of people any more. According to the tourism data by National Tourism Administration from 2015 to 2017, the average number of tourists in China increased by more than 10% per year, and exceeded 5 billion person-time till 2017. At the same time, with the popularity of social media, many visitors share their experiences on various online travel platforms, such as Mafengwo and Baidulvyou. More and more travel enthusiasts, especially young people, are willing to spend time recording and sharing their memories and experiences in their journey via travelogues.

Travelogues contain rich information such as photos and texts, which can provide important references for other travel enthusiasts. Unlike the advertising information released by official website of scenic spots and travel agencies, the feelings and experiences shared by travelogues can better fit the needs of travel enthusiasts. However, with the increasing number of tourists, the number of travelogues from social travel apps like Mafengwo is significantly increasing, which may result in information overload. Therefore, how to mine useful travel knowledge from heterogeneous crowdsourced data and make recommendations based on users requests has become an important and practical research problem.

Though there have been numerous studies [1] [2] [3] [4] on tour information recommendation, but there are still several open issues to be addressed. First, the recommendations in existing work are usually at a coarse-grained level, they recommend routes consisting of multiple scenic spots within a city. For example, the Forbidden City is a famous scenic spot in China. It covers a wide area and involves numerous attractions such as Hall of Central Harmony and Hall of Supreme Harmony. Visiting such scenic spots often needs several hours to a full day. Hence, fine-grained route recommendation over these attractions become a necessary. Second, existing work mainly makes use of textual information, hence the recommendation is mostly text-based without any visual information. However, travelogues contain rich multimedia information such as photos, where the photos can be leveraged to provide visual information about an attraction in a scenic spot. Third, existing studies rarely consider travelling contexts of users. The contexts such as the season, time (daytime or evening), and companions will impact the final selection of travel routes.

It is non-trivial to address the issues above as we need to tackle two main challenges. The first challenge is how to identify the cross-modal semantic relationship between the text description of an attraction and the relevant photos in travelogues. Photos can provide visual and intuitive information about an attraction in a scenic spot to complement text description. However, due to the diversity of writing style and the lack of labels for the embedded photos, it is difficult to match the photos of attractions with their textual description in travelogues. Assuming that the first challenge is solved, a large number of photos would be associated with each attraction in a scenic spot. Therefore, the second challenge is how to select an optimal subset of photos to visually represent the attractions of the scenic spot.

We address these challenges by developing CrowdTravel, a context-based, visual-augmented approach for fine-grained tour information recommendation. Compared to existing studies, CrowdTravel has several unique features: i) it provides fine-grained travel route recommendation, i.e., at the attraction level; ii) it provides visual and textual tour information; iii) it tailors to diverse tourist-specific contexts.

To sum up, our main contributions are as follows:

- We present a hybrid CNN-RNN model to identify the relationship between photos of attractions and their textual description from travelogues in various forms.
- We propose CrowdRank, a crowd-intelligence-based photo selection method for context-driven, diversity-representativeness-balanced scenic spot photo selection. Specifically, CrowdRank first establishes inter-photo links based on the features extracted by the CNN-RNN model

to build the so-called scenic graph, then it ranks photos and selects representative photos in each cluster.
- We evaluate the effectiveness of CrowdTravel based on a real dataset collected from Baidulvyou and Mafengwo, containing 11542 travelogues and 11228 travel albums of eight popular scenic spots in China. The results indicate that CrowdTravel outperforms the baseline methods.

II. RELATED WORKS

A. Tourism Knowledge Mining

In general, mining tourism knowledge from massive social network data can be divided into two aspects: popular scenic spot mining and travel route mining.

Popular scenic spot mining. Ji et al. [5] explored the city's popular scenic spots by building graph model based on user-community-contributed multimedia data. However, they did not consider different user preferences. Several works considered user historical travel information. For example, Yu et al. [6] recommended personalized tour package to help users make travel plans based on user preference of locations. Cheng et al. [7] proposed a probabilistic Bayesian learning framework to recommend personalized scenic spots, which considers both users traveling history and geo-tagged photos on social networks. Memon et al. [8] presented an expert tourist guide, which considers not only user's travel history, but also travel time. Kurashima et al. [9] proposed personalized recommendation via the Bayesian network techniques based on users behavior and users' relationships. Jiang et al. [10] calculated similarities of user preference to predict user's favorite scenic spots. Moreover, considering that user preferences change over time, Wang et al. [11] recommended similar scenic spots based on the scenic spot queries provided by users. Cao et al. [12] further clustered a large amount of geo-tagged photos based on geo-location, then similar scenic spots can be matched with the user queries based on text or images.

Travel route mining. Yuan et al. [13] [14] extracted frequent sequence relationships between multiple tourist attractions from a large number of travel blogs through frequent pattern mining. Choudhury et al. [15] proposed a recursive greedy algorithm to generate a travel route which takes travel time into consideration based on user-uploaded photos with time stamps, latitude and longitude on Flicker. Kennedy and Naaman et al. [16] adopted a tag-location-vision strategy to group city views, which are represented by geo-tagged Flickr photos. However, these works just present some coarse-grained travel routes, which only include the name and sequence of scenic spots. Moreover, the visual results of the final mining are also single-dimensional (e.g., text), which cannot meet the different needs of users. To overcome the shortcomings of macro-level travel route recommendation, there has been some research that recommends detailed paths at a micro-level in the form of attraction sequence in recent years. For example, Orellana et al. [17] mined the path of tourists in the Dutch National Park from peoples GPS trajectories. However, this method requires accurate GPS data which is difficult to obtain in the scenic area. Guo et al. [18] utilized labeled photos embedded in travelogues to recommend visual travel routes.

B. Cross-modal Data Mining

Cross-modal data (such as photo, video, text, speech) has gradually become mainstream data in many real-world scenarios. Therefore, how to mine and model the cross-modal multimedia data has been an important research problem. In general, there are two main traditional methods to solve the cross-modal data mining problem, including the feature extraction-based method [19] and Canonical Correlation Analysis-based method [20]. For example, Gao et al. [19] utilize the feature extraction-based method to extract photo and context features from the tagged photo data in order to provide relevance of searched photos. In addition, several works have been done based on the method of Canonical Correlation Analysis [21] [22] [23], which learns a subspace to maximize the correlation among data of different media types.

Recently, the rapid development of deep neural networks brings new opportunities to the multimodal representation. For example, Srivastava et al. [24] used the deep Boltzmann machine to generate as many descriptions as possible for the photos, associating the photos with the texts. Similarly, Karpathy et al. [25] used a multimodal recurrent neural network to annotate pictures. Ngiam et al. [26] proposed a multi-modal deep learning (MDL) method that combined audio and video into an automatic encoder that improved the classification of speech signals for noise input and learned the joint representation of cross modalities. The method of deep learning multimodal information by using neural network can maximize the correlation between different types of media data, and is suitable for the text information connection with photo. In this paper, we leverage a combination of Convolutional and Recurrent Networks (CNN-RNN) to associate the photos with the texts.

C. Diversity-oriented Photo Ranking

Different from traditional graph ranking, diversity graph ranking focuses on how to balance the representativeness and diversity of nodes so that the selected Top-k nodes set better covers the entire network. Diversity graph ranking has been mainly based on three types of methods: maximizing marginal benefit, competing random walk, and reinforcement between clustering and ranking.

The method of maximizing the marginal benefits first defines a benefit function that measures the benefits of nodes, and then selects the nodes that can maximize the benefit function one by one. For example, Zhai et al. [27] first proposed a sub-topic coverage method to improve the ranking diversity. Zhang et al. [28] proposed the Affinity Graph Model, which reordered topics by using diversity and information richness. Tong et al. [29] used a personalized PageRank algorithm to measure the centrality of a single node. The diversity of node sets is measured by the edge weights. The other kind of methods is based on random walk. Most of the diversity graph ranking work based on random walk introduces

the competition mechanism between nodes in the random walk process, and realizes the node diversity by enabling the connected nodes to compete with each other. For example, DivRank proposed by Mei et al. [30] provided a random walk strategy with varying transition probability over time and introduces the Rich-get-richer mechanism in random walk. The GRASSHOPPER algorithm proposed by Zhu et al. [31], GSparse [32] proposed by Saule et al., and the manifold ordering algorithm MRSP [33] proposed by Cheng et al, are all methods in which nodes are sequentially selected using an Absorbed Random Walk with Absorbed States. For the selected node, set it to the absorbing state and then random walk and then select the next node.

The third kind of diversified graph ranking exploits the reinforcement between ranking and clustering. Cluster-dependent ranking and ranking-based clustering are iteratively updated to form a diversified ranking. RankClus [34] and NetClus [35] mutual enhancement by ranking and clustering simultaneously improve ranking and clustering. In this paper, we apply an extension of the reinforcement between clustering and ranking method to balance the representativeness and diversity of nodes, which realizes diversity by clustering and utilizes the crowd intelligence of the travelers to infer photo representativeness within each cluster.

III. PROBLEM STATEMENT AND SYSTEM OVERVIEW

A. Problem Formulation

The aim of CrowdTravel is to achieve context-based visual recommendation of scenic spots. The input of CrowdTravel is $U = \{T, A, N, C\}$, a collection of travelogues T, albums A, attractions N, and users input context C, related to the same scenic spot. $T = \{I_t, D_t\}$ represents the travelogues collected form Mafengwo, where I_t denotes the photos and D_t means the texts that describe photos in the travelogues. $A = \{I_a, D_a\}$ represents the albums collected form Baidulvyou, where I_a denotes the photos and D_a means the texts that describe photos in the albums. Besides, $N = \{N_1, N_2 \cdots N_n\}$ denotes the name of all the attractions within the scenic spot, and C denotes users input context. The output of CrowdTravel is a set of diverse yet representative photos regarding the attractions of each scenic spot, which are presented in an order of the recommended travel route.

B. System overview

Figure 1 presents an overview of the CrowdTravel framework, consisting of three major modules: 1) Cross-modal travel information matching, which identifies attraction photos from travelogues by using visual and textual semantic learning. 2) Crowd-intelligence-based tourism photo selection. It firstly builds inference models to learn the contextual cues of associated photos, which is then used for irrelevant photo filtering. Then, a graph model is established according to the association between the text and the photo in each set of data. Based on the graph, a photo selection method that can choose photos that satisfy diversity and representativeness is proposed.

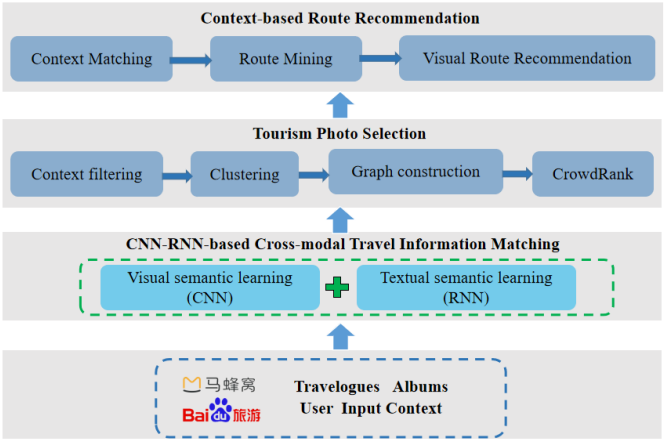


Fig. 1. The Framework of CrowdTravel

3) Context-based visual route recommendation, which mines the associated route (a sequence of attractions) according to different contexts input provided by the user and uses the selected photos to visualize the route mined from travelogues of the scenic spot.

IV. THE DESIGN OF CROWDTRAVEL

We now introduce the framework of CrowdTravel, including cross-modal travel information matching, tourism photo selection, and context-based route recommendation.

A. Cross-Modal Travel Information Matching

To provide visual travel route recommendation, we first need to identify the relationship between textual description about attractions and the relevant photos from the travelogues. Considering the fact that the photos and text descriptions of the same attraction are similar, the travel information matching problem can be considered as a multi-classification problem since there are many attractions at the same scenic spot. However, the different attractions at the same scenic spot are sometimes similar, resulting in small inter-class variance among multiple attractions, as is illustrated in the example of the three attractions in Fig. 2. Meanwhile, user-contributed photos are often taken from different angles and views, so there could be a significant intra-class variance of the same attraction, as is shown in Fig. 2. The small inter-class variance and large intra-class variance have made it difficult to classify different attractions within the same scenic spot. Considering that deep learning can learn deep semantic representations and extract fine-grained features, which have been proved an effective way for photo classification. We propose the CNN-RNN model that combines visual information and natural language description to learn latent semantic representations, as shown in Fig. 3, CNN-RNN first leverages advanced deep learning models (e.g., ResNet-50 [36]) to extract visual features and textual features respectively, which we call them the visual semantic learning and textual semantic learning modules. After

Fig. 2. Example of intra-class variance and inter-class variance

the visual and textual latent feature representations are learned respectively, they are jointly used to form the final multi-modal feature representation. Based on the multi-modal feature representation, the softmax layer is leveraged to get the final classification result.

Visual semantic learning. The visual semantic learning extracts the visual features from original photos, which focuses on the texture, color, and even the semantic parts. CNN model has good performance in visual classification, which has been proved by a large number of recognition tasks, such as texture recognition and fine-grained photo classification [37] [38]. In order to efficiently extract visual features, we use ResNet50 that is pre-trained on a subset of the ImageNet database. On top of the last layer of ResNet50 network, we add a fully connected layer to adjust the dimension of final visual feature representation to p. As a result, the model can learn rich feature representations from a wide range of photos. We then fine-tune the pre-trained ResNet50 on our photo dataset. We take the original photo I as the inputs of the ResNet50 to obtain the feature f_v, which is the result of the latent feature representations for photos. Denoting p dimensional visual feature representation as $f_v \in R^p$, the result of the last layer in the visual semantic learning can be represented as:

$$f_v = \sigma(W_v \cdot f_R) \tag{1}$$

where $\sigma(\cdot)$ is the ReLU activation function, f_R is the visual feature representation obtained from pre-trained ResNet50 and W_v is the weight matrix of the fully connected layer in the visual semantic learning.

Textual semantic learning. The textual semantic learning extracts the textual features from the text descriptions of photos. This module utilizes the cross-modal analysis to learn the correlation between the natural language descriptions and the textual features, and provides a flexible and compact way to match cross-modal information. Text descriptions of photos is a sequence of words, and there might be dependencies between them. To learn and use long-term dependencies to classify sequence data, we employ a word-level LSTM neural network. An LSTM network is a type of recurrent neural network (RNN) that can learn long-term dependencies between time steps of sequence data. We first take the original text descriptions D as the input of Word Embedding, each word in the text is represented as a word embedding vector. The embedding vector for each word is initialized with the pre-trained word embedding on the given dataset. For each text, the corresponding k dimensional word embedding vector is denoted as $x \in R^k$. On top of the last layer of LSTM network, we add a fully connected layer to adjust the final textual feature representation to p. We take x as the inputs of the LSTM model to obtain feature f_t, which is the result of the textual feature representations for texts. Denoting p dimensional textual feature representation as $f_t \in R^p$, the result of the textual semantic learning can be represented as:

$$f_t = \sigma(W_t \cdot f_t) \tag{2}$$

where $\sigma(\cdot)$ is the ReLU activation function, h_t is the textual feature representation obtained from LSTM and W_t is the weight matrix of the fully connected layer in the textual semantic learning. In order to maximize the correlation between a photo and its corresponding description as well as minimize correlation with texts from other classes, we apply the joint training method. During the joint training process with the textual semantic learning, the parameters of pre-trained ResNet50 neural network are kept static to avoid overfitting. The visual feature representation f_v and textual feature representation f_t will be concatenated to form the multi-modal feature representation $f \in R^{2p}$ denoted as:

$$f = f_v \bigoplus f_t \tag{3}$$

Given training data $\left\{\left(x^{(1)}, y^{(1)}\right), \ldots \left(x^{(m)}, y^{(m)}\right)\right\}$, $x^{(i)} = \{I_i, D_i\}$, i=1,2,N in which I_i denotes photos, D_i denotes the corresponding textual description and $y^{(i)} \in Y = 1, 2, , k$ indicates class label. The classifier function $F : f \rightarrow Y$ is learned by minimizing the loss function in Eq.(4):

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \cdot log\left(p(y^{(i)} = j|x^{(i)}; \theta)\right)] \tag{4}$$

where m is the number of photo-text pairs in the training set and k is the number of classes.

B. Graph Model-based Tourism Photo Selection

Based on photo-text association identification, we have plentiful labeled photos for different attractions. In order to give users a better visualization of recommendation, well-chosen photo should be shown. The challenge of tourism photo selection is making the balance of diversity and representativeness. Considering that the *WithWhom*, *Month* and *DayorNight* will impact the recommendation result provided to users, we need filter travelogues according to user input context before

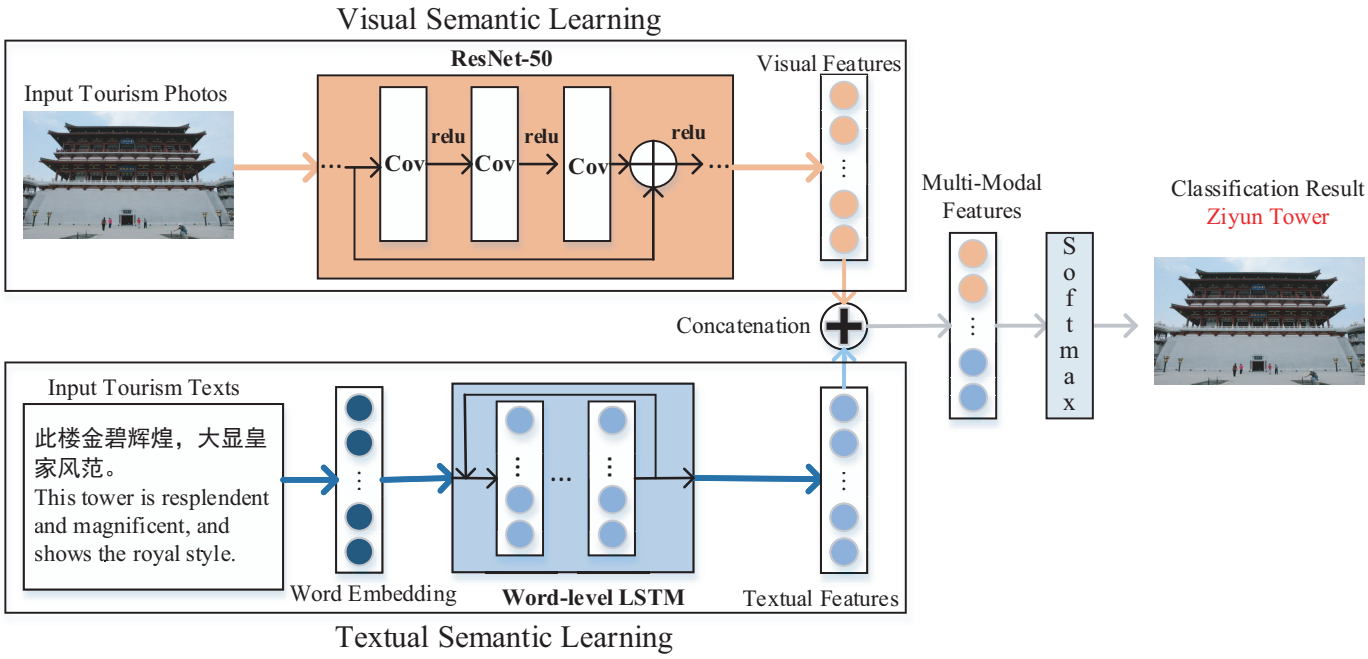


Fig. 3. Overview of our CNN-RNN model

selecting diverse and representative photos. We first extract the Month based on the date of the travelogue, and filter travelogues taken in the same season according to user inputs. The *WithWhom*, and *DayorNight* information is then extracted from each travelogue.

In order to present attraction photos taken in different travel time to users, we train a CNN model to identify the DayorNight of photos embedded in the filtered travelogues. With the relevant photos after context filtering, we need to select a subset of them for visual attraction presentation that can consider both data diversity and representativeness. Diversity means the final selected photo sets cover different angles, while representativeness means each photo in the final selected photo sets is of high quality and the view is shared by most people. The inputs of tourism photo selection algorithm include the photo set $I = I_1, I_2, , I_n$ from each scene, the visual features vector f_v generated by CNN over all photos, and the textual feature vector f_t generated by RNN over all photo descriptions, where n is the size of the photo set. The out is a selected set of diverse and representative photos O, the size of which is predefined according to the requirements. To ensure output diversity, the photos are first grouped into r clusters by the k-means clustering algorithm, where r equal to the size of O. We then construct a scenic graph based on content and visual link among the photo-text pairs within each cluster. Finally, the most representative photos in each cluster are selected to form O using CrowdRank. Overall, both photo clustering and photo representativeness are used to achieve a balance between visual diversity and representativeness.

Scenic Graph Construction. We construct a Scene Graph for diversity and representativeness calculation within each scene. Assuming there are n photos within a given scene. A Scene Graph is defined as an undirected graph with n nodes, with each node representing a photo. Two nodes are assigned a link if the similarity of the deep photos semantic vector f_v and the deep text semantic vector f_t ranges among top 50% of all nodes. Correspondingly, when establishing the edge between nodes, the similarity between the nodes is first calculated using equation (5):

$$W_p(i,j) = W_c(i,j) \times W_t(i,j) \tag{5}$$

Where $W_c(i,j)$ denotes the Cosine similarity of the photos of the i^{th} and j^{th} nodes using their visual feature vectors extracted by CNN, and $W_t\,(i,j)$ represents the Cosine similarity of the text of the i^{th} and j^{th} nodes using their textual features vectors extracted by RNN. These two factors are multiplied to generate the overall correlation $W_p\,(i,j)$ between the i^{th} and j^{th} nodes.

CrowdRank. Inspired by PageRank [39], we present the CrowdRank algorithm to infer photo representativeness within each scene. Different from PageRank, our graphs are undirected and the edges are weighted. Based on the Scene Graph construction, photos are analogous to the pages, and context and content similarities are modeled as links. Our algorithm is based on two factual assumptions(1) A photo is more likely to be representative for a scene if the photo is associated with more similar photos or linked with more similar descriptors. (2) A photo is more likely to be representative for a scene if it is contributed by a high-level user in the travel online platform. Initially, the representativeness of each node is assigned a uniform value $\frac{1}{n}$, where n is the total number of undirected graph nodes. The iteration of each undirected graph is the same as the principle of PageRank. The representativeness of

each node is calculated by equation (6).

$$R_i = \sum_{j=1, j\neq i}^{n} \frac{W_p(i,j) \times W_a(i,j)}{c_i} \times R_j \tag{6}$$

subject to:

$$\sum_{i=1}^{n} R_i = 1 \tag{7}$$

where R_i is the representativeness of the ith photo in the scene graph, c_i is the total number of connections between the ith photo and other photos, $W_a\,(i)$ is the user level of the author of i^{th} photo. In each iteration, the popularity of each photo is updated based on the links with other photos. The representativeness of all photos is normalized after each iteration subject to constraint (4). This representativeness is calculated iteratively on the Scenic Graph to get the representativeness of each photo within the current scene. Similar to PageRank, the shrinkage of CrowdRank can be guaranteed through the principle eigenvector of W. Finally, we get r preferred photos in r clusters.

C. Context-based travel route recommendation

After the photo-text association identification and photo selection, multiple diverse and representative photos have been generated for each attraction. This module aims to recommend a personalized travel route based on the user context. The travel route is the attraction sequences that tourists visited in the scenic spot. In the travelogues, photos and contexts are usually organized in the same order as the users travel route. Therefore, refer to [18], we can extract a travel route of a travelogue as a candidate route for recommendation.

V. EXPERIMENTS

A. Datasets

We use two datasets for evaluation. We use the name of the scenic spots as a search term to crawl the information form Mafengwo and Baidulvyou. Both datasets contain 8 popular scenic spots in China, namely Forbidden City, Summer Palace, West Lake, Huangshan Mountain, Tang Paradise, Mountain Resort, Gulangyu and Jiuzhai Valley. Finally, after pre-processing, we have crawled over 16,000 attraction photos together with their textual descriptions, and the average number for each attraction is approximately 2,000.

B. Cross-Modal Travel Information Matching

We evaluate our CNN-RNN model by comparing to the following baselines:

- **Traditional Machine Learning (TML).** In this approach, SURF features are first extracted from photos and TF-IDF features are extracted from the corresponding text descriptions. Traditional machine learning classification methods, such as SVM, Naive Bayes, Random Forest, are then used to identify attraction photos.
- **Visual Semantic Learning (VSL).** Only CNN is used to learn deep visual features to classify attraction photos on photo data.
- **Textual Semantic Learning (TSL).** Only RNN is used to learn deep textual features to classify attraction photos on text data.
- **CVL [40].** This method first uses CNN and RNN to calculate the probability of each category respectively, and then adds the two probabilities to each category. Finally, the model discovers the category with the highest probability as the final result.

TABLE I
PERFORMANCE OF TRAVEL INFORMATION MATCHING

Method	TML	VSL	TSL	CVL	CNN-RNN
Accuracy(%)	70	75	72	78	80

We use accuracy to evaluate the performance of different methods, the results are illustrated in Table 1.We can see that the proposed CNN-RNN method significantly outperforms to other baselines. This demonstrates that visual information and textual descriptions complement with each other in fine-grained photo classification.

C. Graph Model-based Tourism Photo Selection

In this subsection, we evaluate our Clustering&CrowdRank method by comparing to the following baselines:

- **Clustering.** This method uses the Cosine distance of the Bag-of-Visual-Words. The central photos within each cluster are considered as the most representative photos of this cluster.
- **SimilarityRank.** This method uses the Cosine distance of the features extracted by the Bag-of-Visual-Words and the TF-IDF feature within each scene cluster. Cosine distance is used to represent the overall feature associations.
- **PhotoRank [5].** Within each scene cluster, PhotoRank uses the Cosine distance of the deep visual features (extracted by CNN) and deep textual features (extracted by RNN). Cosine distance is adopted as the overall feature associations.

Photo selection aims to make a trade-off between diversity and representativeness. We conduct a user study to evaluate the effectiveness of the proposed method. As shown in Fig. 4, the user study result indicates that Clustering, SimilarityRank, PhotoRank and Clustering&CrowdRank receive an average support of 15%, 19%, 31% and 35% respectively.

D. Context-based travel route recommendation

We use the Tang Paradise as an example to present our recommended route. For Tang Paradise, we assume that the context information specified by the user is the elderly, October, daytime, which means a tour with the elderly in the day in October. The recommended visual path is shown in Fig.5 (a). In order to verify the effectiveness of our method, we then change the context as children, May, evening. In this case, the result is shown in Fig.5 (b). Compared to Fig.5 (b), the unique recommendation attractions for the elderly in Fig.5 (a) are Tang Bazaar and Luyu Teahouse. The former contains many antiques, calligraphy, and paintings, and is suitable for

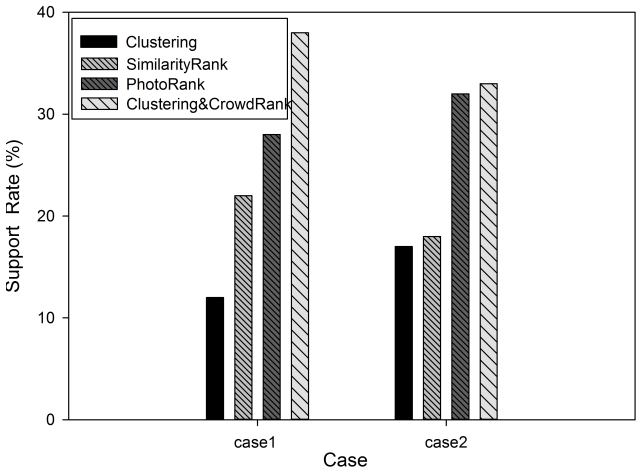
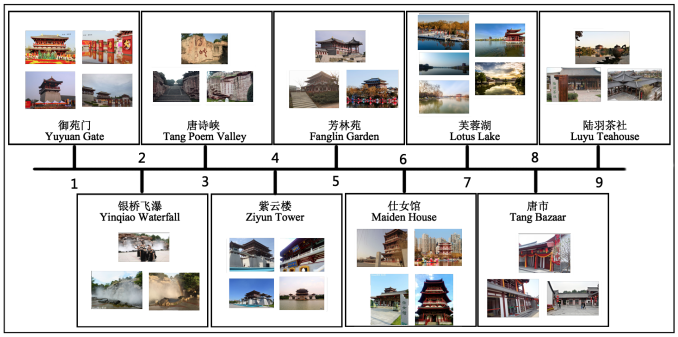


Fig. 4. Tourism photo selection performance



(a) The route for elder people in the daytime in October

(b) The route for taking children in the evening in May

Fig. 5. Recommendation route for two different contexts to Tang Paradise

the elderly to enjoy. The latter provides a recreational area for elder people to have a tea break while they are tired. For children, Fig.5 (b) recommends the Water Screen Film and Fengmingjiutian Theater. These shows are more attractive to children. It can be seen from the recommendation path that different travel groups have different preferences in different travel time and the final result varies in age and time. The selected photos can characterize the attraction in different aspects, giving the users a full understanding of the attraction. The options of daytime and evening situations lead to more different results, helping people better plan their route and travel time. Tang Paradise is a relatively niche scenic spot, which lacks ofcial information, and we can leverage the rich crowdsourced data to demonstrate the practical effect of CrowdTravel.

VI. CONCLUSION

In this paper, we present CrowdTravel which leverages the rich crowdsourced data in travelogues for visual attraction route recommendation. A combination of visual and textual semantic learning is used to identify photo-text association. Combining visual and textual information is effective in association identification and up to 80% attraction photos can be matched correctly. Further, a photo ranking approach is proposed for visual exhibition that satisfies both diversity and representativeness. Our system is evaluated by using crawled travelogues datasets from travel websites. The results show that our method outperforms the baseline methods. Finally, the user study results indicate that the photo selection approach is promising, which further validates the potential of our framework in advanced travel exploration and related applications. In future works, we will consider more context and expert knowledge when recommending our final results to real travelers.

ACKNOWLEDGEMENT

This work was partially supported by the National Key R&D Program of China(2017YFB1001800), the National Natural Science Foundation of China (No. 61772428,61725205).

REFERENCES

[1] M. Nomiyama, T. Takeuchi, H. Onimaru, T. Tanikawa, T. Narumi, and M. Hirose, "Xnavi: Travel planning system based on experience flows," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, p. 27, 2018.

[2] C.-H. Hsu, C.-L. Ku, Y.-J. Chang, Y.-S. Wang, U.-D. Trân, W.-H. Cheng, C.-Y. Yang, C.-Y. Hsieh, and C.-C. Lin, "itour: Making tourist maps gps-enabled," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, p. 139, 2018.

[3] M. Nomiyama, T. Takeuchi, H. Onimaru, T. Tanikawa, T. Narumi, and M. Hirose, "Travel planning system considering experience flows based on driving histories," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 177–180, ACM, 2016.

[4] W.-M. Pang, "Context-aware scene recommendation for travel photography," in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pp. 17–20, ACM, 2015.

[5] R. Ji, X. Xie, H. Yao, and W.-Y. Ma, "Mining city landmarks from blogs by graph modeling," in *Proceedings of the 17th ACM international conference on Multimedia*, pp. 105–114, ACM, 2009.

[6] Z. Yu, H. Xu, Z. Yang, and B. Guo, "Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 151–158, 2016.

[7] A.-J. Cheng, Y.-Y. Chen, Y.-T. Huang, W. H. Hsu, and H.-Y. M. Liao, "Personalized travel recommendation by mining people attributes from community-contributed photos," in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 83–92, ACM, 2011.

[8] I. Memon, L. Chen, A. Majid, M. Lv, I. Hussain, and G. Chen, "Travel recommendation using geo-tagged photos in social media for tourist," *Wireless Personal Communications*, vol. 80, no. 4, pp. 1347–1362, 2015.

[9] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, "Travel route recommendation using geotags in photo sharing sites," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 579–588, ACM, 2010.

[10] K. Jiang, P. Wang, and N. Yu, "Contextrank: personalized tourism recommendation by exploiting context information of geotagged web photos," in *2011 Sixth International Conference on Image and Graphics*, pp. 931–937, IEEE, 2011.

[11] Y. Wang and L. Cao, "Discovering latent clusters from geotagged beach images," in *Advances in Multimedia Modeling*, pp. 133–142, Springer, 2013.

[12] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, and T. S. Huang, "Aworldwide tourism recommendation system based on geotaggedweb photos," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2274–2277, IEEE, 2010.

[13] H. Yuan, H. Xu, Y. Qian, and K. Ye, "Towards summarizing popular information from massive tourism blogs," in *2014 IEEE International Conference on Data Mining Workshop*, pp. 409–416, IEEE, 2014.

[14] H. Yuan, H. Xu, Y. Qian, and Y. Li, "Make your travel smarter: Summarizing urban tourism information from massive blog data," *International Journal of Information Management*, vol. 36, no. 6, pp. 1306–1319, 2016.

[15] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu, "Automatic construction of travel itineraries using social breadcrumbs," in *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pp. 35–44, ACM, 2010.

[16] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How flickr helps us make sense of the world: context and content in community-contributed media collections," in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 631–640, ACM, 2007.

[17] D. Orellana, A. K. Bregt, A. Ligtenberg, and M. Wachowicz, "Exploring visitor movement patterns in natural recreational areas," *Tourism Management*, vol. 33, no. 3, pp. 672–682, 2012.

[18] T. Guo, B. Guo, J. Zhang, Z. Yu, and X. Zhou, "Crowdtravel: Leveraging heterogeneous crowdsourced data for scenic spot profiling and recommendation," in *Pacific Rim Conference on Multimedia*, pp. 617–628, Springer, 2016.

[19] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 363–376, 2013.

[20] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*, pp. 162–190, Springer, 1992.

[21] H. Bredin and G. Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2, pp. II–233, IEEE, 2007.

[22] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[23] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4437–4446, 2015.

[24] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, pp. 2222–2230, 2012.

[25] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

[27] C. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval," in *ACM SIGIR Forum*, vol. 49, pp. 2–9, ACM, 2015.

[28] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma, "Improving web search results using affinity graph," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 504–511, ACM, 2005.

[29] H. Tong, J. He, Z. Wen, R. Konuru, and C.-Y. Lin, "Diversified ranking on large graphs: an optimization viewpoint," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1028–1036, ACM, 2011.

[30] Q. Mei, J. Guo, and D. Radev, "Divrank: the interplay of prestige and diversity in information networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1009–1018, Acm, 2010.

[31] X. Zhu, A. Goldberg, J. Van Gael, and D. Andrzejewski, "Improving diversity in ranking using absorbing random walks," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 97–104, 2007.

[32] O. Küçüktunç, E. Saule, K. Kaya, and Ü. V. Çatalyürek, "Diversifying citation recommendations," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 4, p. 55, 2015.

[33] X.-Q. Cheng, P. Du, J. Guo, X. Zhu, and Y. Chen, "Ranking on data manifold with sink points," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 177–191, 2013.

[34] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: integrating clustering with ranking for heterogeneous information network analysis," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 565–576, ACM, 2009.

[35] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 797–806, ACM, 2009.

[36] T. Akiba, S. Suzuki, and K. Fukuda, "Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes," *arXiv preprint arXiv:1711.04325*, 2017.

[37] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

[38] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*, pp. 647–655, 2014.

[39] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[40] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5994–6002, 2017.