EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

A travel route recommendation algorithm based on interest theme and distance matching

Xi Cheng

Correspondence: 20030032@sasu.edu.cn
Sichuan University of Arts and Science, Dazhou 635000, China

Abstract

To solve the problem of low accuracy of traditional travel route recommendation algorithm, a travel route recommendation algorithm based on interest theme and distance matching is proposed in this paper. Firstly, the real historical travel footprints of users are obtained through analysis. Then, the user's preferences of interest theme and distance matching are proposed based on the user's stay in each scenic spot. Finally, the optimal travel route calculation method is designed under the given travel time limit, starting point, and end point. Experiments on the real data set of the Flickr social network showed that the proposed algorithm has a higher accuracy rate and recall rate, compared with the traditional algorithm that only considers the interest theme and the algorithm which only considers the distance matching.

Keywords: Travel route recommendation, Interest theme, User interest, Distance matching

1 Introduction

In recent years, the research of recommender system has developed rapidly. Various recommendation systems are also widely used in e-commerce, social networking sites, e-tourism, Internet advertising, and many other fields, and these recommendation systems show superior effects and prospects [1–3]. With the rise of more and more online travel websites (such as Expedia, Travelzoo, tuniu), more and more online data can describe users' interests and preferences. This makes tourism product recommendation become one of the hotspots of recommendation system research [4, 5].

At present, many mature recommendation algorithms have been widely used in traditional product recommendation, such as collaborative filtering algorithm [6], content-based recommendation algorithm [7], and hybrid recommendation algorithm [8]. However, a large number of existing studies show that tourism product recommendation is very different from traditional film product recommendation [9–11], and the differences are as follows.

First of all, users usually do not often or batch purchase tourism products, which leads to the sparse correlation matrix of "user products." Secondly, the description of tourism product information is diverse and complex. Small parameter changes will lead to completely different tourism products, such as scenic spots and schedule, hotel, and vehicle selection. However, this kind of tourism product with inherent relevance points to the common interests of users. Third, users often do not pay attention to tourism products for a long time, that is, they leave visit records on e-tourism websites, and often after browsing objectives and arrangements. Then, they began to browse tourism products, which led to a large number of cold start users in online travel data. Therefore, the traditional recommendation algorithm is difficult to apply to tourism recommendations. Generally speaking, the abovementioned recommendation technologies solve the problem of travel recommendation to a certain extent. However, this recommendation technology is only suitable for tourism data with relatively simple data structure, or relies on geographic information data, so it is difficult to fully capture users' real-time interest preferences.

The data used in this paper is the real web server log of tourism enterprises, which contains rich tourism product information and a large number of user behavior click records. The recommendation engine can be built to accurately capture users' interests based on their real-time click stream. Then, personalized tourism products are recommended to users according to their interests. In order to improve the accuracy of the recommendation system, a travel path recommendation algorithm based on interest and distance matching is proposed. The core idea is to calculate the user's preference topic and acceptable distance according to each user's travel history data and add them as weights into the recommendation model to get a new personalized travel path recommendation algorithm.

The rest of the paper is organized as follows. We review the related work in the first section. The second section gives the definition and algorithm of necessary concepts. The third section gives the experimental results. Finally, the fourth part draws a conclusion.

2 The related work

With the continuous development of social networks, users' social network information is increasing. How to effectively mine valuable information from social network information plays an irreplaceable role in the development of social network [12]. In social networks, users can upload text information, location information, and time information and share this information with friends and nearby people. At present, more and more scholars recognize the importance of social network information and devote themselves to the research of social network information mining.

The idea of social network data mining is similar to GPS trajectory data mining. In GPS trajectory data mining, the main applications include association rules, abnormal behavior, travel mode, and GPS trajectory recommendation [13]. The data acquisition time is strictly limited by equal time intervals, which is reflected in shahed [14]. In social network trajectory data mining, applications mainly include location recommendation, path recommendation, and behavior preference recommendation [15]. Data collection time is discrete and random, which is the main difference between social network trajectory data and GPS trajectory data.

At present, there are many processing methods of social network data mining, including clustering, classification, and other traditional technologies. Among them, the clustering method is used to discover group pattern mining methods in social networks, and it has a good effect in recommending user path and location. MapReduce [16] framework is widely used in the large-scale data processing. At present, the method of combining clustering algorithm with MapReduce framework for big data analysis and processing is gradually developed, such as DBSCAN clustering algorithm based on MapReduce, which has achieved good results.

The mining methods of group pattern mainly include group, group, escort, assembly, and platoon. The literature [17] introduces different mining methods of group patterns in detail. Swarm is a group pattern mining technology with weak time constraints. It only needs to satisfy the condition that the number of different tracks appearing at the same point is greater than the set threshold. Although team formation and escort have more time constraints than team formation, this strong constraint will also lead to the decline of accuracy. The row model is described in [18]. The platoon model combines the advantages of the above group models and adapts to different applications by allowing control of continuous-time constraints.

Personalized recommendation methods mainly include content-based recommendation, collaborative filtering recommendation, association rule-based recommendation, utility-based recommendation, knowledge-based recommendation, and combination recommendation [19]. At the same time, there are many recommendation strategies, and different recommendation strategies produce different recommendation results. However, in the personalized travel route recommendation based on group pattern, due to the lack of semantic information, the traditional group pattern mining leads to an incomplete personalized recommendation.

3 Algorithm implementation

3.1 Basic definition

Given a directed weighted graph $G = (E, V)$, V is the set of nodes, and E is the set of edges. As shown in Fig. 1, a node $p \in V$ represents a *POI*. Each POI_p has category

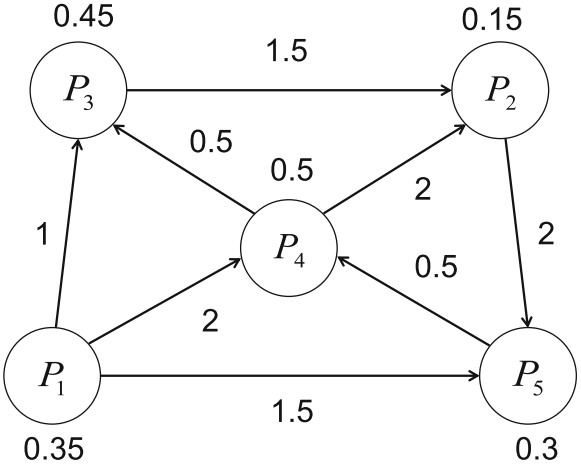

Fig. 1 An example of a route recommendation

attributes Cat_p (such as church, museum, beach), longitude, and latitude, and the value on the node p represents the score of POI_p. C represents a collection of categories of all *POIs*. In the attribute (c_i, D_i) of the node p_i, c_i represents the category attribute of the *POI*, and D_i represents the distance of the *POI*. Each directed edge (p_i, p_j) represents a feasible route between two *POIs*, the number of edges is $|E|$, and the weight on the edge represents the travel time (in h) of the continuous access to the two *POIs*.

A travel route is a sequence consisting of multiple travel POIs, denoted as $R = \{p_1, p_2, \ldots, p_N\}$, where p_i is the tourist location included in the route, and N is the number of locations.

The travel time between two *POIs*. The travel time required by the user from POI_{P_x} to POI_{P_y} can be defined as follows:

$$T^{Travel}\left(p_x, p_y\right) = Dist\left(p_x, p_y\right)/speed \tag{1}$$

where $Dist(p_x, p_y)$ represents the distance between p_x and p_y, which is calculated by the Haversine formula [20]. Suppose the user walks to play and takes a speed of 6 km/h.

The preference vector of a user u is expressed as $IntP(u) = \langle Int(u, c_1), Int(u, c_2), \ldots, Int(u, c_i)\rangle$, where $Int(u, c_i)$ represents the degree of preference of the user u for the *POI* category c_i.

Given a user u and the *POI* collection he/she has been to define his/her historical travel footsteps in chronological order $S_u = ((p_1, t^a_{p_1}, t^d_{p_1}), (p_2, t^a_{p_2}, t^d_{p_2}) \ldots, (p_n, t^a_{p_n}, t^d_{p_n}))$. Each triplet $(p_x, t^a_{p_x}, t^d_{p_x})$ consists of a POI_{P_x} that the user has visited, a time $t^a_{p_x}$ that reaches p_x, and a time $t^d_{p_x}$ that leaves p_x, consisting of three elements. The first photo taken by the user in each *POI* is the time of the user's arrival and the last photo is the time of the user's departure. The user's access time in p_x (that is, the user u's stay in p_x) can be valued by the difference between $t^a_{p_x}$ and $t^d_{p_x}$. Similarly, for the travel sequence S_u, $t^a_{p_1}$ and $t^d_{p_n}$ represent the start and end times of the journey, respectively. For simplicity, this paper represents $S_u = ((p_1, t^a_{p_1}, t^d_{p_1}), (p_2, t^a_{p_2}, t^d_{p_2}) \ldots, (p_n, t^a_{p_n}, t^d_{p_n}))$ as $S_u = (p_1, p_2, \ldots, p_n)$.

The travel footprint $S_u = (p_1, p_2, \ldots, p_n)$ of the user u is known, and if $t^a_{p_{x+1}} - t^d_{p_x} > \tau$, S_u is divided into a number of individual travel sequences (that is, sub-sequences of S_u). In other words, if the time between consecutive accesses of two *POIs* is greater than the threshold τ, the travel footprint is divided into a number of different tourist subsequences. The time threshold value τ is selected in this document as 8h.

This paper gives a *POI* scoring method considering location distance and user preferences. The score for POI_{p_i} is expressed as $score(p_i)$:

$$score(p_i) = \alpha \times Int(u, c_i) + (1-\alpha)D(p_i) \tag{2}$$

where c_i is the category of POI_{p_i}, $D(p_i)$ is the distance of POI_{p_i}, and α is the user adjustment parameter, which is used to adjust the proportion of user interest preference and *POI* distance in the route.

3.2 Tourist route recommendation framework

As shown in Fig. 2, the travel route recommendation in this paper is divided into the construction of the *POI* association graph and the learning of the user's interest preference, as well as the route recommendation. The construction of the *POI* association graph and the learning of the user's interest preference are performed offline, and the distance of the *POI* and the user's interest preference can be obtained by analyzing the photos taken by the user. The route recommendation is online, assuming that the user wants to go to the city with *mPOIs*, recorded as $P = \{p_1, p_2, \ldots, p_m\}$. According to the *POI* set *P*, time budget *B*, starting point POI_{p_1}, and ending point POI_{p_n}, the route with the highest score is recommended to users by using the proposed algorithm which combines user interest preferences and POI distance based on the orientation problem.

3.3 Construction of POI correlation graph

The construction of the POI association diagram takes place offline. In this paper, POI in the tourism sequence of all users is used as the node in the graph, which represents the tourism place, and the continuous access of users in the tourism sequence generates the edges in the graph.

The structure of the photo data shared by the user is (PhotoID, UserID, time, longitude, latitude, category). From this structure, the photo data contains the exact space-time location information of the user. Based on the longitude and latitude of each photo, the Haversine formula [20] is used to calculate the distance between each photo shared by the user and each POI in the city visited. If the distance is less than 200m,

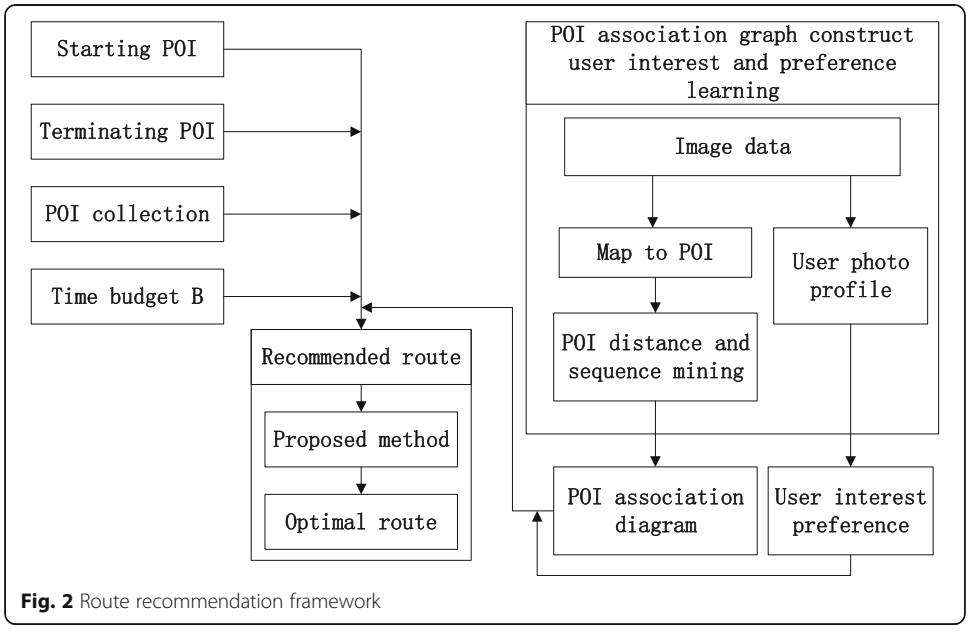

Fig. 2 Route recommendation framework

the photograph is considered to be taken at POI, so as to obtain the list of users' POI $S_u = (p_1, p_2, \ldots, p_n)$.

A time-based user interest preference is presented by using the user's historical travel footprint. When a user goes to a POI to play, he/she will stay at the POI for a certain period of time. The access time (i.e., the stay time) of each POI that each user has visited is calculated from the historical travel footprint of all users according to definition 4, so that the average time required for any user to access any one of the POIs can be calculated. In the travel route recommendation in this article, $\overline{V}(p)$ is used to indicate the average access time of his/her POI_p for any user. The average access time required for each POI_p is as follows:

$$\overline{V}(p) = \frac{1}{n}\sum_{u\in U}\sum_{p_x\in S_u}\left(t^d_{p_x} - t^a_{p_x}\right)\sigma(p_x = p); \forall p \in P \tag{3}$$

where U represents all users, and n represents the number of users accessing p in U. $\sigma(p_x = p) = \begin{cases} 1, p_x = p \\ 0, \text{other} \end{cases}$.

However, the average access time of the user at each POI does not truly reflect his/her degree of interest preference for this type of POI. Therefore, a time-based user interest preference is proposed in this paper. The preference degree of user u to category attribute c of POI is calculated from the following equation.

$$Int(u, c) = \sum_{p_x\in S_u}\frac{\left(t^d_{p_x} - t^a_{p_x}\right)}{\overline{V}(p_x)}\sigma\left(Cat_{p_x} = c\right); \forall c \in C \tag{4}$$

where Cat_p represents the category attribute of $POIp, \sigma(Cat_{p_x} = c) = \begin{cases} 1, Cat_{p_x} = c \\ 0, other \end{cases}$. The above equation determines the interest of user u in the category attribute c for a particular POI. Relative to the average access time of all users in the same POI, it is calculated based on the time spent by users in each POI with category attribute$c$. In other words, a user may spend more time accessing the type of POI he or she is interested in, which in turn determines the user's level of interest in such POI.

3.4 Proposed algorithm

Orienteering problem (OP) is a directional problem, which is described as follows. In a directed weighted graph G(V,E), V is the set of all points on the graph, and E is the set of all edges on the graph. Each point has its score (score, which can be expressed as gain), and each edge has its weight (weight, which is the walking time between two points). The start and end points are specified. Select partial points from G. Then, plan a path through the selected points, the starting points, and ending points. At the same time, under the premise of not exceeding a certain time budget, the total weight score of the path is maximized.

OP has been widely used in travel route recommendations. The route recommendation algorithm is proposed in consideration of POI distance and user interest based on the orientation problem. Based on the set P, the time budget B, the starting point

POI_{p_1}, and the end point POI_{p_n}, the proposed algorithm recommends a route $R = \{p_1, p_2, \ldots, p_n\}$ that satisfies the time budget B and has the highest score. The time budget is calculated by function $Cost(p_x, p_y)$, $Cost(p_x, p_y) = T^{Travel}(p_x, p_y) + \overline{V}(p_y)$. Therefore, the travel route recommendation model in this paper can satisfy the integer programming problem with multiple constraints, which is expressed as follows:

$$Max \sum_{i=2}^{N-1} \sum_{j=2}^{N} x_{i,j} score(p_i) \tag{5}$$

where $x_{i,\ j} = 1$ indicates the route from i to j, which goes through the edge (p_i, p_j), otherwise $x_{i,\ j} = 0$. The above equation satisfies the following constraints:

$$\sum_{j=2}^{N} x_{1,j} = \sum_{i=1}^{N-1} x_{i,N} = 1 \tag{6}$$

$$\sum_{j=2}^{N} x_{k,j} = \sum_{i=1}^{N-1} x_{i,k} \le 1; \forall k = 2, 3, \ldots, N-1 \tag{7}$$

$$\sum_{i=1}^{N-1} \sum_{j=2}^{N} Cost(i, j) x_{i,j} \le B \tag{8}$$

$$2 \le u_i \le N \tag{9}$$

$$u_i - u_j + 1 \le (N-1)(1-x); \forall i, j = 2, 3, \ldots, N \tag{10}$$

Equation (5) is an objective function that maximizes the POI distance and user interest preferences in the recommended route. Eqs. (6) to (10) are the constraints of Eq. (5). Equation (6) ensures that the user's travel starting point is p_1 and the end point is p_n. Equation (7) ensures that the user's travel route is coherent and that each POI in the route has only been visited once. Equation (8) guarantees that the time spent by the user on the entire trip is within budget B. Assuming that u_x is the location of $POIx$ in route R, Eq. (9) and Eq. (10) ensure that there are no sub-patrol routes in the integer programming problem proposed in this paper.

4 Experimental results and analysis

4.1 Experimental data and the data preprocess

This article uses the Flickr dataset of web and mobile data management (WAMDM), which contains two aspects. One is 319,110 photos of New York City, including photo ID (identity), photo location (latitude and longitude), photo time, photo tag information, and user ID. The other is the address book of 12,991 Flickr users. In this experiment, we use the leap one out cross-validation method [21] used in most

recommender system research to verify the algorithm. It circulates one data in the whole data set as the test set and the other as the training set and calculates various evaluation indexes according to the prediction results of each cycle. In order to facilitate the experiment, the data are processed as follows: (1) delete users who take less than 5 photos per poi and (2) delete users who have only visited one or two POIs. After data preprocessing, the final experimental data set consists of 55,451 images taken by users and 165,830 images of Geotag.

4.2 Evaluation index

The accuracy of recommendation is the most important index of the evaluation algorithm. In this paper, precision and recall are used as criteria for measuring the pros and cons of the algorithm, which are expressed as follows.

$$precision = |P_r \cap P_v| / |P_r| \tag{11}$$

$$recall = |P_r \cap P_v| / |P_v| \tag{12}$$

Precision represents the probability that the user is interested in the recommended route, and recall represents the probability that a user's favorite POI is recommended. The higher the accuracy and recall, the better the recommended effect. P_r represents the POI set in the recommended route, and P_v represents the POI set visited by users in the real tourism sequence.

4.3 Experiment result

In order to verify the effectiveness of the algorithm, this paper compares the traditional travel path recommendation algorithm. In the traditional travel route recommendation algorithm, POI distance and user interest preference are considered as the criteria. Under different time budgets B, the traditional algorithm is compared with the algorithm proposed in this paper, and a travel route recommendation algorithm based on POI distance and user interest preference is proposed. The experimental results are shown in Figs. 3 and 4.

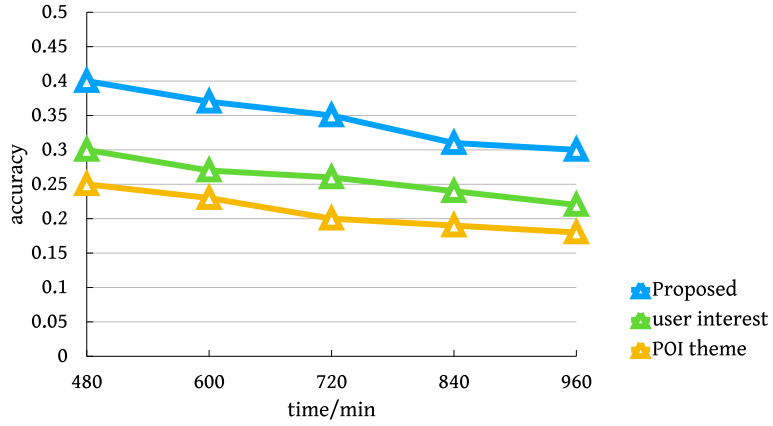

Fig. 3 Results of accuracy

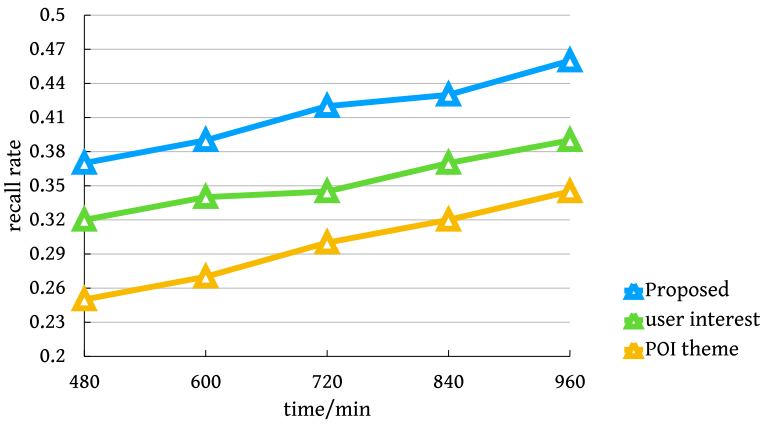

Fig. 4 Results of recall rate

In terms of accuracy, it can be seen from Fig. 3 that the proposed algorithm has higher accuracy than the traditional algorithm only considering user interests and poi topics. In terms of recall rate, as shown in Fig. 4, this algorithm has a higher recall rate than the traditional algorithm which only considers user interests and poi topics. One of the influencing factors is that the proposed algorithm and the algorithm only considering the user's interest consider the user's interest, because users prefer to visit the places they are interested in. The high accuracy and recall rate of the algorithm show that the algorithm can more accurately recommend the path reflecting the real travel sequence of users.

5 Conclusion

Based on the positioning problem, this paper establishes a travel recommendation model and proposes a personalized travel route recommendation algorithm. The algorithm comprehensively considers POI distance and user interest preference, recommends the most suitable route to users, and realizes the tourism route recommendation framework by using Flickr photo set with geographical tags. Finally, the experimental results show that the proposed algorithm has a higher recommendation accuracy and recall rate than the traditional algorithms which only consider POI distance or user interest preference. The next step is to further study the intelligent optimization algorithm to solve the orientation problem, so as to improve the efficiency of the algorithm and reduce the cost.

Abbreviations

GPS: Global Position System; DBSCAN: Density-based spatial clustering of applications with noise; POI: Point of interest; WAMDM: Web and mobile data management

Acknowledgements

TangSeng Huang helped perform the analysis with constructive discussions.

Author's contributions

XI cheng, as the primary contributor, completed the analysis, experiments, and paper writing. The author read and approved the final manuscript.

Funding

Research on the Construction of Health Informationization of People in the old Revolutionary Base Area - Dazhou City(SLQ2020SB-015).

Availability of data and materials
The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Received: 10 May 2021 Accepted: 30 June 2021
Published online: 03 August 2021

References

1. M.F. Alhamid, M. Rawashdeh, H. Dong, et al., Exploring latent preferences for context-aware personalized recommendation systems [J]. IEEE Transactions on Human-Machine Systems **46**(4), 615–623 (2016)
2. A. Klašnja-Milićević, M. Ivanović, B. Vesin, et al., Enhancing e-learning systems with personalized recommendation based on collaborative tagging techniques [J]. Appl Intell **48**(6), 1519–1535 (2018)
3. D. Ayata, Y. Yaslan, M.E. Kamasak, Emotion based music recommendation system using wearable physiological sensors [J]. IEEE Trans Consum Electron **64**(2), 196–203 (2018)
4. R. Colomo-Palacios, F.J. García-Peñalvo, V. Stantchev, et al., Towards a social and context-aware mobile recommendation system for tourism [J]. Pervasive and Mobile Computing **38**, 505–515 (2017)
5. C. Benfares, Y.E.B. El Idrissi, A. Amine, *Smart city: recommendation of personalized services in patrimony tourism [C]//2016 4th IEEE International Colloquium on Information Science and Technology (CiSt). IEEE* (2016), pp. 835–840
6. J. Chen, H. Zhang, X. He, et al., *Attentive collaborative filtering: multimedia recommendation with item and component-level attention [C]//International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM* (2017), pp. 335–344
7. L. Cui, L. Dong, X. Fu, et al., A video recommendation algorithm based on the combination of video content and social network [J]. Concurrency and Computation: Practice and Experience **29**(14), e3900 (2017)
8. Y. HUANG, Y. Jinxin, S.U.N. Wei, Research of hybrid recommendation algorithm based on improved bipartite network and expert trust [J]. Value Engineering **36**(19), 160–164 (2017)
9. Q. Liu, E. Chen, H. Xiong, et al., A cocktail approach for travel package recommendation [J]. IEEE Transactions on Knowledge & Data Engineering **26**(2), 278–293 (2013)
10. C. Tan, Q. Liu, E. Chen, et al., Object-oriented travel package recommendation [J]. ACM Trans Intell Syst Technol **5**(3), 1–26 (2014)
11. C.M. Lee, J.J. Thomas, *Travel route recommendation based on geotagged photo metadata [C]//International Visual Informatics Conference* (Springer, Cham, 2017), pp. 297–308
12. L.Q. Nie, X.M. Song, T.S. Chua, in *Proc. of the Synthesis Lectures on Information Concepts, Retrieval, and Services*. Learning from multiple social networks (Morgan & Claypool Publishers, 2016)
13. P.F. Yin, M. Ye, W.C. Lee, Z.H. Li, in *Proc. of the 18th Pacific-Asia Conf. on PAKDD*. Mining GPS data for trajectory recommendation (Springer-Verlag, 2014), pp. 50–61
14. A. Eldawy, M.F. Mokbel, S. Al-Harthi, A. Alzaidy, K. Tarek, S. Ghani, in *Proc. of the ICDE*. SHAHED: a MapReduce-based system for querying and visualizing spatio-temporal satellite data (2015), pp. 1585–1596
15. Y. Shen, L.G. Zhao, J. Fan, Analysis and visualization for hot spot based route recommendation using short-dated taxi GPS traces. Information **6**(2), 134–151 (2015)
16. Y.B. He, H.Y. Tan, W.M. Luo, S.Z. Feng, J.P. Fan, MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data. Frontiers of Computer Science **8**(1), 83–99 (2014)
17. Y.X. Li, J. Bailey, L. Kulik, Efficient mining of platoon patterns in trajectory databases. Data Knowl Eng **100**, 167–187 (2015)
18. Q. Fan, D.X. Zhang, H.Y. Wu, K.L. Tan, A general and parallel platform for mining co-movement patterns over large-scale trajectories. PVLDB **10**(4), 313–324 (2016)
19. T. Hasuike, H. Katagiri, H. Tsubaki, H. Tsuda, in *Proc. of the SMC*. A route recommendation system for sightseeing with network optimization and conditional probability (2015), pp. 2672–2677
20. D. Gavalas, C. Konstantopoulos, K. Mastakas, et al., Review: mobile recommender systems in tourism [J]. J Netw Comput Appl **39**(1), 319–333 (2014)
21. C. Xin, C. Gao, C.S. Jensen, Mining significant semantic locations from GPS data [M]. VLDB Endowment (2010)

Publisher's Note