Personalized Tourism Route Recommendation System Based on Dynamic Clustering of User Groups

Weiwei Yin, Yan Sun*, Jing Zhao
Panjin Vocational and Technical College
Panjin, Liaoning, China
e-mail: zhaojing@pjvtc.edu.cn

2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC) | 978-1-7281-9018-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/IPEC51340.2021.9421158

***Abstract*—Tourism path dynamic planning is an asynchronous group model planning problem. It is required to find group patterns with similar trajectory behavior under the constraint of unequal time intervals. Traditional trajectory group pattern mining algorithms often deal with GPS data with fixed time interval sampling constraints, so they can not be directly used in coterie pattern mining. At the same time, traditional group pattern mining has the problem of lack of semantic information, which reduces the integrity and accuracy of personalized travel route recommendation. Therefore, this paper proposes a semantic based distance sensitive recommendation strategy. In order to efficiently process large-scale social network trajectory data, this paper uses MapReduce programming model with optimized clustering to mine coterie group patterns. The experimental results show that: under MapReduce programming model, coterie group pattern mining with optimized clustering and semantic information is superior to traditional group mode in personalized travel route recommendation quality, and can effectively process large-scale social network trajectory data**

Keywords—MapReduce, Data Relevance, Big Data, Tourism Route Recommendation

I. INTRODUCTION

As a popular social application of picture sharing, Instagram is widely used by tourists to record travel information such as location, time, uGc(user generated context), which mainly includes travel route, travel density distribution, preferences, movement mode and so on [1-2]. Therefore, how to effectively mine large-scale Instagram track data plays an extremely important role in tourist route recommendation, but the existing track group pattern mining is only applicable to GPS track data sampled at equal time intervals [3]. In addition, with the continuous development of social networks, the scale of data has gradually increased, forming track big data, and MapReduce, as a parallel programming framework, provides convenience for large-scale data processing [4]. It can be seen that the social network trajectory big data currently faces the following problems [5-6]: 1) whether the clustering algorithm can improve the efficiency through MapReduce parallel optimization processing; 2) the traditional trackgroup pattern has the input data sampling constraint of equal interval, and whether it can find a discrete and random method for instagram data trackgroup pattern mining. 3) At present, the existing group model only considers trajectory information, but does not consider the impact of social network UGC information on travel route recommendation. Can UGC be combined with social network trajectory information to complete personalized travel route recommendation based on trajectory group mode. These problems have challenges in the tourism route recommendation, so far, it has not been found that the combination of the above points has been carried out.

II. RELATED WORK

With the continuous development of social network, 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. In social network, users can upload text information, location information and time information, and share these information with friends and nearby people [7-8]. Nowadays, more and more scholars are aware of the importance of social network information, and have been involved in the research of social network information mining.

The idea of social network data mining is similar to that of GPS trajectory data mining. In GPS trajectory data mining, the main applications include association rules, abnormal behavior, travel mode and GPS trajectory recommendation. Data collection time has strict equal time interval limit, which is reflected by SHAHED. In social network trajectory data mining, the application mainly includes location recommendation, route recommendation and behavior preference recommendation. The 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 in social network data mining, including clustering, classification and other traditional technologies. Among them, the clustering method is used to find out the group pattern mining method in social network, which has a good effect on recommending user route and location. In large-scale data processing, MapReduce framework is widely used. At present, the method of combining clustering algorithm with MapReduce framework for big data analysis and processing is gradually developed. For example, DBSCAN clustering algorithm based on MapReduce has achieved good results.

Group pattern mining methods mainly include swarm, flock, harmony, gathering, platoon, etc. Swarm is a group pattern mining technology with weak time axis constraint. It only needs to satisfy the condition that different trajectories

appear at the same place at the same time more than the threshold. While flock and convoy have stronger time constraints than swarm, but this strong constraint also leads to the decrease of accuracy. The platform mode integrates the advantages of the above group modes and adapts to different applications by allowing control of continuous time constraints. The paper introduces the platoon group mode in detail.

Personalized recommendation methods include content-based recommendation, collaborative filtering recommendation, association rule-based recommendation, utility based recommendation, knowledge-based recommendation and combination recommendation. At the same time, there are many kinds of recommendation strategies, and different recommendation strategies produce different recommendation results. However, in the personalized travel route recommendation based on group pattern, the traditional group pattern mining is not perfect due to the lack of semantic information.

III. Optimized Clustering by NMR-DBSCAN

At present, most of the researches on tourist route recommendation are implemented in a single machine environment, which cannot face the explosive growth of data volume. Therefore, MapReduce parallel programming framework based on Hadoop platform is applied to social network trajectory big data processing to improve algorithm efficiency. Because DBSCAN algorithm can find the advantages of arbitrary shape clustering in noisy spatial data, this paper combines DBSCAN algorithm with MapReduce parallel framework for clustering processing. In order to prevent the data information from being lost during data partition processing, boundary points will be stored in two adjacent partitions at the same time. At the same time, in order to balance the load of nodes and improve the efficiency of the algorithm, the Prbp algorithm is improved to reduce the boundary points. In addition, the node memory overflow will lead to the failure of the entire MapReduce task, and the improved nprbp algorithm can ensure that the memory of each node is below the threshold, and has the minimum number of boundary points under the fixed partition conditions.

The nprbp algorithm mainly consists of three steps [9]: (1) building a grid with the width of 2eps, because DBSCAN clustering algorithm takes EPS as the radius, which can save enough clustering information. (2) Calculate the total number of element points in each piece and the accumulated amount of element points; (3) select the best fragment, traverse each partition, find the least element point partition which is closest to the threshold of the set node element, and iterate. Figure 1 sets the partition number n = 2, the fourth partition element is the least, and the two partition elements meet the threshold. Therefore, 4 is the partition boundary. Fig. 1 (a) is the partition processing graph, and Fig. 1 (b) and l (c) are the partition results.

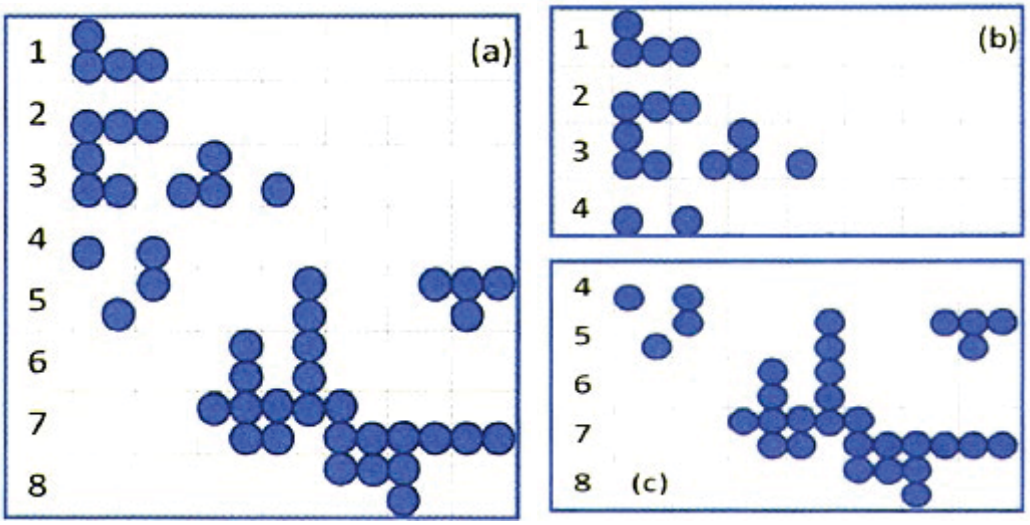

Figure 1 Partition map of nPRBP algorithm

See algorithm 1 for pseudo code of nMR-DBSCAN clustering processing. the input is trajectory data set and the output is clustering result. Line 1 ~ line 13 program performs nPRBP partition processing on track data. Lines 14-24 complete the DBSCAN clustering processing based on MapReduce. In the merging stage, different partition clusters with the same core boundary points are found for merging processing. In the relabeling stage, the merging results and partition clustering results are relabeled to obtain the final clustering results. After the clustering processing is finished, the user tracks are expressed in the form of clustering sequence in time sequence, for example, $Tra(o_i)=\{C_1, C_2,...,C_k\}$. The leaker set can be obtained from definition 2, for example: $Tra(o_2)=\{C_l, C_2\}$, minc=3, then the track object o_2 is not leaker.

Algorithm 1.nMR-DBSCAN algorithm

. Input: tra: trajectory data;

Var tra=readinput data;

{Step 1: running nPRBP ontra}

1. S=buildSliceUse2Eps(tra.Eps); /*initializing slices for each dimension*/

2. p=PRBP(tra); /*running PRBP on tra; */

3. P.add(p);

4. For eachpartitionslicepin P do

5. For each slice S inSdo

6. If5.total>p.totaland J.index>p.index do

7. If sliceNumber<n and j.total<size do

8. P.add(s);

9. Delete p from P:

10. Endif

11. EndIf

12. Endfor

13. EndFor

{Step II: running DBSCAN in MapReduce phase}

14. For eachp artition p in P do /*selectpartitionin Map phase*/

15. DBSCANClustering(p); /*running DBSCAN on p in Reduce phase*/

16. For eachpointPtsinpdo

17. If Pts.islnnerdo /*storing result of inner points to local file*/

l 8. Output(partition.index, Pts.index+Pts.id;

19. Endif

20. Else /*storingresult ofboundary pointsto HDFS*/

21. writeFile(partition.index.Pts.index+Pts.id+ Pts.isCore);

22. End else

23. Endfor

24. Endfor

IV. Pattern Mining of Coterie Trajectory Group

Because the traditional ObjectGrowth algorithm for mining track group patterns is inefficient in finding track closed patterns, the ClusterGrowth algorithm is proposed to mine closed coteries patterns. Fig. 2 is an object trajectory and cluster diagram under eps=2 (neighborhood radius) and Minpts=2 (neighborhood point threshold).

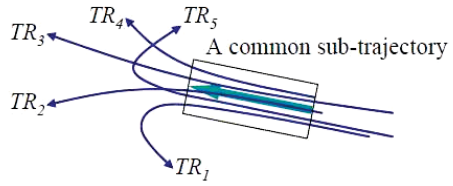

Figure 2 Object trajectory and clustering

In addition to the two pruning strategies, the verification rules are used to verify whether the coterie is a closed coterie. The idea of test rule is similar to that of posteriori pruning strategy, because posteriori pruning can only judge pruning conditions between parent node and child node, while inspection rule prevents pruning omission among other nodes. As shown in Figure 3, the orange nodes are marked as closed courtiers because they do not meet the prior pruning and posteriori pruning conditions, but the orange nodes can be pruned by the inspection rules, and the gray nodes in Figure 3 are the final closed courtiers.

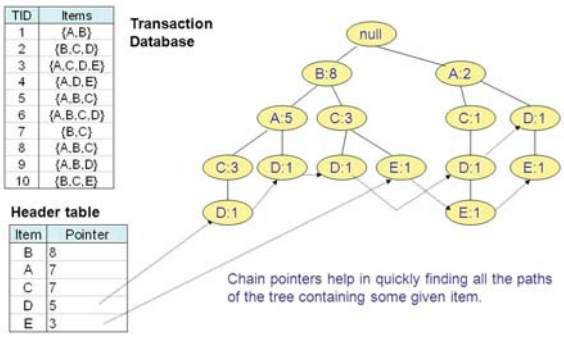

Figure 3 Cluster growth algorithm

V. Personalized Travel Route Recommendation Based on Semantics

Traditional group pattern mining has the problem of missing semantic information, which reduces the quality of personalized travel route recommendation. Therefore, two semantic based recommendation strategies DRSS and CRSs are proposed, and TF IDF technology combined with cosine similarity is used to measure semantic correlation.

Tf-idf is a commonly used weighting technology for information retrieval and data mining. The main idea is that if a word or phrase appears frequently in one article and rarely appears in other articles, it is considered that the word or phrase has good classification ability. In a given document, word frequency (tf) refers to the frequency of a given word in the document. For words in a particular document, the importance is expressed as [10]:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

Among them, $tf_{i,j}$ denote the importance of word frequency of word i in file j, $n_{i,j}$ denote the number of times the word i appears in file j, and $\sum_k n_{k,j}$ represents the sum of the occurrence times of all words in file j.

The word IDF can be measured by the total number of documents and the number of words appearing in the corpus. Meanwhile, the word IDF can be expressed as follows:

$$idf_i = log \frac{|D|}{1+\left|\left\{j : t_i \in d_j\right\}\right|} \quad (2)$$

Among them, idf_i is the reverse file frequency of the word i, |D| is the total number of files in the corpus, and $j : t_i \in d_j$ is the number of files containing the word i. From formula (1) and formula (2), tf-idf is obtained as follows:

$$tf\text{-}idf_{i,j} = tf_{i,j} \times idf_{i,j} \quad (3)$$

Formula (3) shows that high word frequency and low file frequency can produce tf-idf with high weight, that is, tf-idf can filter out common words and keep important

words.

Cosine similarity generates word frequency vectors from words generated by tf-idf, and then obtains cosine similarity from vector cosine angle calculation formula:

$$cos\theta = \frac{\sum_{i=1}^{n}(A_i - B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (4)$$

The closer the cosine value is to 1, the more similar the two vectors are. The semantic similarity is calculated by combining tf-idf technology with cosine similarity.

Distance plays a very important role in the choice of tourist routes. Tourists will try their best to find the shortest route without affecting the travel location. Therefore, the combination of semantic information and distance information can improve the quality of personalized recommendation, which is called DRSS.

Firstly, semantic information and distance information of each track are extracted from closed courtiers. The distance factor is inversely proportional to the recommendation coefficient. The longer the distance is, the less likely the tourists will choose this route. The greater the correlation between input semantics and track semantics, the more likely they are to be selected by tourists, that is, the semantic correlation is directly proportional to the recommendation coefficient.

In group based personalized travel route recommendation, conformity plays an important role in personalized travel route recommendation. When tourists come to a strange city, the tourism plan will be more inclined to the tourist route with more tourists. At the same time, both the conformity factor and semantic relevance are proportional to the recommendation coefficient. Therefore, a recommendation strategy based on semantic and conformity is proposed, which is called CRSs. According to the idea of PageRank, the score of each adjacent cluster between the starting point and the destination is calculated. The score between adjacent clusters is affected by the sequence of adjacent clusters and the interest degree included in the cluster.

VI. Conclusion

In order to improve the efficiency of searching for closed coteries, cluster growth algorithm is used in the stage of pattern mining of coterie group. A priori pruning and a posteriori pruning are used to compress the search space. Finally, a verification rule is proposed to find the closed courtiers. The experimental results show that the efficiency of clustergrowth algorithm is higher than that of objectgrowth algorithm.

In the stage of personalized travel route recommendation based on group pattern, the traditional group pattern mining leads to the imperfection of personalized recommendation due to the lack of semantic information. Therefore, two recommendation strategies based on semantic DRSS and CRSs are proposed to complete the recommendation of personalized travel routes. The experiment shows the tourism routes recommended by DRSS and CRSs.

References

[1] Xiao Chunjing, Xia Kewen, Qiao Yongwei. Tourism Route Recommendation Based on Dynamic Clustering. Computer Applications, 2017, 37 (008): 2395-2400

[2] Lu Wei, Ni Yuhua. Clustering Algorithm of Tourism Routes Based on Equidistant Encryption and Case-based Reasoning. Computer Engineering and Applications, 2010 (11): 223-225

[3] Chen Jianke, Chen Pinghua. Tourism Route Recommendation Algorithm Based on Interest Hotspot Map. Computer Engineering and Design, 2018, 39 (09): 249-254

[4] Zhang Yinghui, Li Xue. Tourism Recommendation Algorithm Based on Fuzzy Clustering. Computer Technology and Development, 2016 (12): 99-102

[5] Yang Fengyi, Ma Yupeng, Bao Hengbin. Mining Self Driving Travel Routes Based on Sparse Trajectory Clustering. Computer Applications, 2020, V.40; No.356 (04): 155-160

[6] Ren Yaopeng. Study on the Classification of Yongji Tourist Population Based on Cluster Analysis. Modern Computer (professional Edition), 2017, 000 (029): 52-55

[7] Zhou Shiping, Zhou Yongwu, Wang Haijuan. Design of Tourism Service Supply Chain Service Package Based on Fuzzy Cluster Analysis Theory. Systems Engineering, 2013, 031 (006): 95-99

[8] Wang Xin, Huang Zhongyi, Wang Xin. Personalized Recommendation Based on - Means Clustering in Network Resources. Journal of Beijing University of Posts and Telecommunications: Self Science Edition, 2014, 21: 69-75

[9] Liu Jiantao. Research on User Polymorphism Clustering in Personalized Recommendation System. Data Analysis and Knowledge Discovery, 2012, 28 (2)

[10] Liu Bing. Two Kinds of Dynamic Cluster Evaluation of Tourism Landform Resources. Journal of Sichuan Agricultural University, 1996, 000 (003): 476-481