Journal of Computer Applications
计算机应用, 2020, 40(4): 1079 - 1084
ISSN 1001-9081
CODEN JYIIDU
2020-04-10
http://www.joca.cn

文章编号:1001-9081(2020)04-1079-06
DOI:10.11772/j.issn.1001-9081.2019081467

基于稀疏轨迹聚类的自驾车旅游路线挖掘

杨奉毅[1,2,3],马玉鹏[1,3*],包恒彬[1,2,3],韩云飞[1,3],马　博[1,2,3]
(1. 中国科学院新疆理化技术研究所,乌鲁木齐 830011; 2. 中国科学院大学,北京 100049;
3. 新疆民族语音语言信息处理实验室,乌鲁木齐 830011)
(*通信作者电子邮箱ypma@ms.xjb.ac.cn)

摘　要:针对自驾车游客加油轨迹稀疏,还原真实旅游路线困难的问题,提出一种基于语义表示的稀疏轨迹聚类算法,用以挖掘流行的自驾车旅游路线。与基于轨迹点匹配的传统轨迹聚类算法不同,该算法考虑不同轨迹点之间的语义关系,学习轨迹的低维向量表示。首先,利用神经网络语言模型学习加油站点的分布式向量表示;然后,取每条轨迹中所有站点向量的平均值作为该轨迹的向量表示;最后,采用经典的k均值算法对轨迹向量进行聚类。最终的可视化结果表明,所提算法有效地挖掘出了两条流行的自驾车旅游线路。

关键词:稀疏轨迹;旅游路线挖掘;轨迹聚类;分布式表示;自驾车旅游
中图分类号:TP391.4　　**文献标志码**:A

Self-driving tour route mining based on sparse trajectory clustering

YANG Fengyi[1,2,3], MA Yupeng[1,3*], BAO Hengbin[1,2,3], HAN Yunfei[1,3], MA Bo[1,2,3]
(1. *The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi Xinjiang* 830011, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*;
3. *Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi Xinjiang* 830011, *China*)

Abstract: Aiming at the difficulty of constructing real tour routes from sparse refueling trajectories of self-driving tourists, a sparse trajectory clustering algorithm based on semantic representation was proposed to mine popular self-driving tour routes. Different from traditional trajectory clustering algorithms based on trajectory point matching, in this algorithm, the semantic relationships between different trajectory points were considered and the low-dimensional vector representation of the trajectory was learned. Firstly, the neural network language model was used to learn the distributed vector representation of the gas stations. Then, the average value of all the station vectors in each trajectory was taken as the vector representation of this trajectory. Finally, the classical k-means algorithm was used to cluster the trajectory vectors. The final visualization results show that the proposed algorithm mines two popular self-driving tour routes effectively.

Key words: sparse trajectory; tour route mining; trajectory clustering; distributed representation; self-driving tour

0　引言

随着国民经济水平提升,私家车保有量迅速增加,自驾车旅游逐渐成为人们旅游出行的热门选择。通过对游客的自驾轨迹进行分析,可以发现流行的自驾车旅游路线,为旅行者的旅行路线的规划提供支持。然而,自驾车游客的活动具有较高的自主性,导致行程轨迹数据难以收集,数据的代表性和覆盖性不足。

本文采用覆盖新疆的加油数据集,挖掘流行的新疆自驾车旅游路线。该数据集记录了用户在新疆的所有加油行为,其中同样包含了所有自驾车游客的加油记录。按照时间先后将加油站点组成序列就可以得到一条加油轨迹,加油轨迹是游客旅游路线的某种采样,能够真实反映游客的时空移动轨迹,可以作为新疆自驾车旅游路线挖掘的重要数据来源。然而,利用加油数据挖掘流行的自驾车旅游路线主要面临两个挑战。首先,原始数据中人员数量众多,加油行为复杂多样,难以准确识别出自驾车游客群体;其次,相比全球定位系统(Global Positioning System, GPS)轨迹数据,加油记录产生的频率非常低,数据十分稀疏,导致同一条加油轨迹中的两个连续轨迹点之间的路径不确定,从中还原出具体的路线十分困难。如图1所示,图中的线条为某游客自驾游的确切轨迹,圆点表示该游客实际的加油地点。可以看出,依靠单一游客稀疏的加油轨迹点数据,并不足以推断出游客实际的旅游路线。

本文的主要工作如下:针对游客群体识别的问题,通过分析已知的游客加油行为,总结出游客加油的基本特征,进而从大量原始加油记录中识别出游客群体。针对轨迹点稀疏问题,受word2vec[1]的启发,提出一种基于语义表示的稀疏轨迹聚类算法:将每个加油站看作一个单词,每条加油轨迹看作一

收稿日期:2019-08-23;**修回日期**:2019-10-19;**录用日期**:2019-11-04。
基金项目:新疆维吾尔自治区自然科学基金资助项目(2019D01A92);新疆天山杰出青年计划项目(2018Q005)。
作者简介:杨奉毅(1994—),男,山东济宁人,硕士研究生,主要研究方向:大数据分析、数据挖掘; 马玉鹏(1979—),男,新疆阜康人,研究员,博士,CCF会员,主要研究方向:物联网、大数据分析; 包恒彬(1995—),男,辽宁本溪人,硕士研究生,主要研究方向:大数据分析、数据挖掘; 韩云飞(1990—),男,山西晋城人,助理研究员,博士,主要研究方向:数据挖掘、计算机视觉; 马博(1984—),男,辽宁鞍山人,副研究员,博士,CCF会员,主要研究方向:大数据分析、知识图谱。

个句子，通过 word2vec 学习加油站点的分布式表示，然后使用每条轨迹中站点向量的平均值表示该加油轨迹，最后通过 k 均值算法完成轨迹聚类，根据聚类结果挖掘流行的自驾车旅游路线。图 2 为本文方法的总体流程。

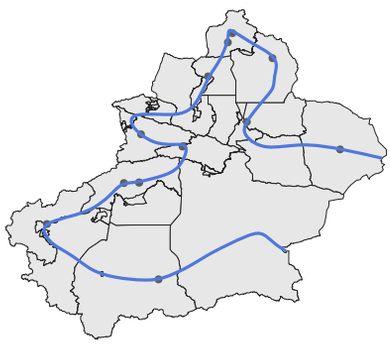


图 1　某游客旅游路线与加油轨迹对比

Fig. 1　Comparison between tour route and refueling trajectory of a tourist

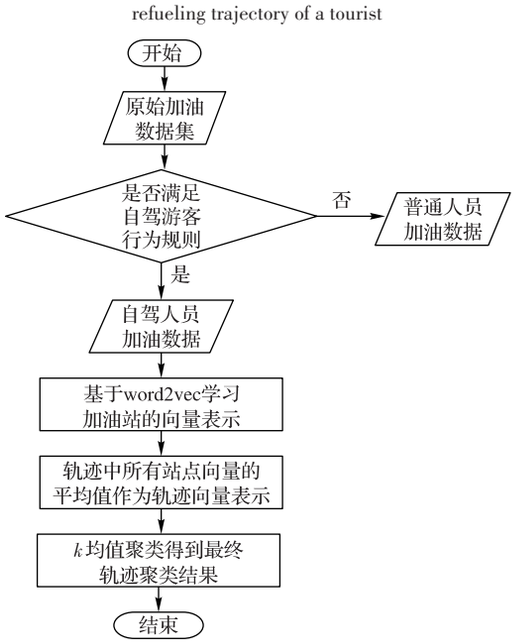


图 2　本文算法的总体流程

Fig. 2　Overall flow of the proposed algorithm

1　相关工作

1.1　旅游路线挖掘

近些年来出现了大量旅游路线规划和推荐的相关研究。目前研究采用的数据主要是用户分享的 GPS 轨迹数据、带地理标签的照片数据和签到数据[2]。Zheng 等[3-5]基于 GPS 轨迹数据做了一系列的旅游路线挖掘和推荐的工作，取得了优秀的成果。Cui 等[6]基于用户的 GPS 轨迹信息，考虑用户的个性化信息，基于协同过滤技术提出了两种旅游路线推荐算法，提高了推荐结果的个性化程度。然而，GPS 轨迹虽然可以反映游客的具体旅行路线，但是相对难以获得。随着基于位置服务的发展，利用社交媒体数据进行旅游路线挖掘和推荐成为新的研究热点。文献[7]提出了一种基于主题模型的协同过滤方法，利用旅游照片进行旅游推荐。文献[8]基于游客签到数据，提出一种基于集体知识的路线推理框架，从不确定轨迹中挖掘流行的旅游路线。目前，采用这些用户分享数据进行自驾车旅游的研究较少，主要原因在于这些数据难以区分普通游客与自驾游客，其中能够完全确定为自驾游客数据的数量很少。文献[9]基于自驾游客分享的 924 条 GPS 轨迹，结合路网信息与旅游景点信息，以季节强度指数、多维缓冲区、核密度等方法分析自驾车游客的时空行为特征。

上述数据大多来自于用户的分享，这些用户仅仅占全部游客的一小部分，因此分析结果带有较大的偏差。与上述研究采用的数据不同，本文采用加油轨迹数据进行自驾车旅游路线的挖掘，数据集本身包含全部自驾车游客在新疆的加油行为，对于分析新疆自驾游的整体情况具有重要作用。

1.2　轨迹聚类

为了发现不同移动对象轨迹中的代表路径或共同趋势，通常将相似的轨迹进行聚类[10]。传统的轨迹聚类方法一般先使用一定的度量方法比较轨迹之间的相似性，之后采用一些经典的聚类算法进行聚类。Lee 等[11]提出了一个轨迹聚类框架，首先将每个轨迹划分为多个子轨迹，然后采用基于密度的聚类方法对这些子轨迹进行聚类。Tang 等[12]提出了出行行为聚类算法，使用结合采样的密度聚类算法解决轨迹数据中的噪声问题。Besse 等[13]定义了一种新的距离度量方法，实现了基于距离的轨迹聚类。

时空轨迹聚类的核心在于衡量轨迹之间的相似性，常用的轨迹相似性度量方法包括动态时间扭曲（Dynamic Time Warping，DTW）[14]、最长公共子序列（Longest Common Subsequence，LCSS）[15]和编辑距离方法 EDR（Edit Distance on Real sequence）[16]。

上述度量方法主要考虑了轨迹点的空间位置信息，适合用于采样频率高的 GPS 数据。然而，游客的加油频率非常低，通常在 2~3 d，加油轨迹十分稀疏，两个连续访问的加油站点之间甚至相隔数百公里。因此，上述方法并不适用于极其稀疏的加油轨迹数据。

1.3　分布式表示

在自然语言处理领域，传统的将单词表示为高维、稀疏向量的方法基本上被基于神经网络的语言模型所取代。神经网络语言模型通过考虑词序和单词的共现来训练，其概念基于分布式假设，即在句子中经常出现在一起的单词具有更高的统计相关性。Mikolov 等[1]提出的 word2vec 是其中的突出代表，它可以简单高效地学习到单词的低维向量表示，在包括机器翻译、情感分析等传统的自然语言处理任务上取得了优秀的表现。

最近，分布式表示的概念逐渐扩展到网页搜索、电子商务、推荐系统等其他领域。研究人员意识到可以将用户的行为序列视作句子，进而学习商品或用户的嵌入表示，如用户的点击、查询或购买序列。分布式表示被用于各种类型的网络推荐中，包括淘宝推荐[17]、求职推荐[18]、应用推荐[19]、房源推荐[20]等。同样，类似的方法也被提出用于社交网络分析，利用网络中节点的随机游走序列学习网络节点的嵌入表示[21]。本文同样利用这一思想，将加油轨迹中的加油站看作单词，整条轨迹看成句子，运用 word2vec 学习加油站的语义向量表示，用于加油轨迹聚类，还原游客的旅行路线。

2　自驾车游客群体识别

原始加油数据集中的记录 $r=\{u,v,s,t,area\}$,其中:u、v、s、t 分别代表司机、车辆、加油站和加油时间戳;$area$ 代表加油站所属的行政区划。从中选取数据构建人员加油轨迹数据集,选取时间范围是2016年1月1日至2018年12月31日。将加油记录按照人员进行重组,每个用户的所有加油记录可以组成一条加油轨迹 $tra=\{s_1,s_2,\cdots,s_n\}$,它是一个用户按时间顺序访问过的加油站点集合。

经过调查可发现自驾车游客通常具有3个特征:1)只来过新疆一次且停留时间不超过30 d;2)具有连续加油行为,相邻两次加油时间间隔小于5 d;3)加油地点不固定且分散在各个地州市,通常不少于3个地州市。同时,为了避免轨迹过于稀疏,本文只对加油次数大于8次的轨迹进行分析。根据以上特征,定义了四条规则,用以自驾车游客群体的识别。对于加油轨迹数据集中的任意一条轨迹 $tra=\{s_1,s_2,\cdots,s_n\}$,如果满足:

$$t(s_n)-t(s_1)\leqslant 30$$

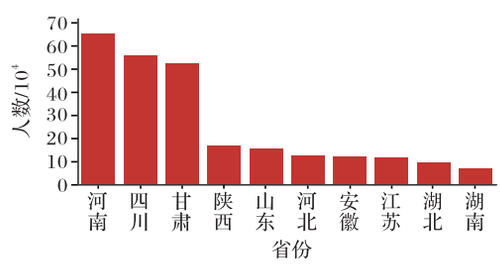
$$t(s_j)-t(s_{j-1})<5;\ 1<j\leqslant n$$

$$different(area)\geqslant 3$$

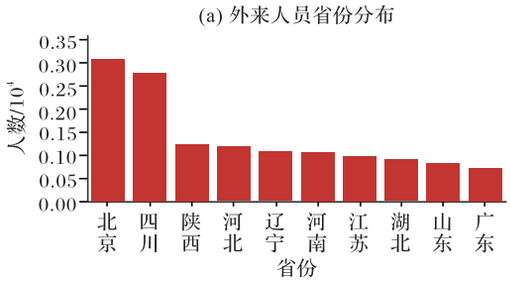
$$n>8$$

则认为轨迹 tra 为自驾加油轨迹,对应人员为自驾游客。

根据以上规则,从加油轨迹数据集中提取出20 646条加油轨迹。为了验证这些轨迹确为自驾游人员所产生,本文对轨迹对应人员的基本特征做了统计分析。首先,这些人员的男女比例高达21∶1,远远高于原始数据集中3.5∶1的男女比例,这符合长途自驾旅游中司机绝大多数为男性的实际情况。其次,游客的主要来源省份与整体情况也有着很大差异,如图3所示。



(a) 外来人员省份分布



(b) 自驾游客省份分布

图3　外来人员与自驾游客的省份分布

Fig. 3　Provincial distribution of migrants and self-driving tourists

图3(a)为原始数据集中外来人员的主要省份分布,这在很大程度上体现了外地人在新疆生活及工作的情况,可以看到,四川、甘肃和陕西等距离较近的省份,与新疆的人员交流比较密切;同时,河南、山东等省份由于人口总量大,在新疆也有大量的人员活动。如图3(b)所示,游客的来源省份发生了很大的变化,较发达省份的人员占比明显增加,来自北京的人员数量最多,而来自甘肃的游客数量很少,这体现了居民收入及消费水平对自驾车旅游的重要影响。最后,对自驾游人员的月份分布做了统计,如图4所示。结果显示,游客的数量受季节变化十分明显,主要集中在7、8、9三个月,与此同时,几乎没有游客选择在冬季进行自驾游,这与新疆的气候条件有着密切的关系,符合新疆旅游市场的基本情况。

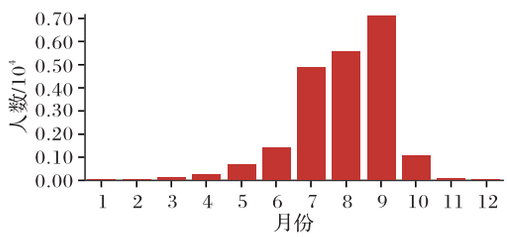


图4　自驾车游客人数随时间变化的统计

Fig. 4　Histogram of self-driving tourist number variation with time

由此可见,本文定义出的规则有效识别出了自驾车游客群体,能够支撑进一步的研究工作。

3　基于语义表示的加油轨迹聚类

3.1　问题定义及分析

定义1　自驾路线。游客实际的自驾路线是一个连续的空间曲线,是游客驾车自驾的确切移动路径。

定义2　加油轨迹 tra。每条加油轨迹都是按照时间先后排序的游客加油的站点序列,是自驾路线的一种采样,形式化表示为 $tra=\{s_1,s_2,\cdots,s_n\}$。加油轨迹 tra 和加油站 s 对应的向量表示分别为 $\boldsymbol{traj}$ 和 $\boldsymbol{st}$。

定义3　加油轨迹聚类。给定游客加油轨迹集 T,我们的目标是将轨迹集 T 划分为 k 个不相交的簇 $C=\{C_1,C_2,\cdots,C_k\}$,同一个轨迹簇的加油轨迹对应同一条自驾路线,$T=\{tra_1,tra_2,\cdots,tra_f\}$。

同一条自驾路线上会有数量众多的加油站,游客的加油行为具有极大的自主性,每条加油轨迹都可以看作是对路线上的所有加油站的随机采样。因此,虽然游客的旅游路线相同,最终的加油轨迹依然有着很大的差别。这些轨迹都反映了旅游路线的部分信息,将这些轨迹进行聚类可以得到完整的旅游路线信息。然而,加油轨迹的高度稀疏性导致基于轨迹点空间相似性的传统轨迹聚类算法无法得到理想的效果,因此本文将自然语言处理中语义相似性的概念引入加油轨迹聚类,用以更好地衡量加油站点之间的相似性。

在自然语言处理领域,语义相似的单词拥有相似的上下文。同样,在加油轨迹中,加油站的上下文通常是同一路线上的加油站,这些站点具有更高的语义相似性。使用word2vec学习站点的语义信息,得到站点的向量表示,之后取所有站点向量的平均值作为加油轨迹的向量,最后使用经典的 k 均值算法实现轨迹聚类。本文算法的整体框架描述如下。

算法1　基于语义表示的加油轨迹聚类。

输入　加油轨迹数据集 T,聚类簇数 k,skip-gram模型窗口大小 m,嵌入向量维度 d。

输出　簇划分 $C=\{C_1,C_2,\cdots,C_k\}$。

初始化权重矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{W}'$

for $tra\in T$ do

SkipGram($\boldsymbol{W}, \boldsymbol{W}', m, d, tra$) //word2vec模型训练

end for

//权重矩阵$\boldsymbol{W}$中的每一行都对应一个站点的向量表示

for $tra \in T$ do

$$\boldsymbol{traj} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{st}_k$$

//轨迹中站点向量的平均值作为轨迹的向量表示

end for

$D = \{\boldsymbol{traj}_1, \boldsymbol{traj}_2, \cdots, \boldsymbol{traj}_f\}$

C = Kmeans(D, k) //k-means算法得到轨迹聚类结果

return C

3.2 站点向量表示

本节将介绍使用word2vec进行站点表示的方法。给定加油轨迹数据集T,目标是学习到每个站点的d维向量表示。实际上,word2vec包括两个模型,分别为连续词袋模型(Continuous Bag-of-Words model, CBOW)和skip-gram模型。其中,CBOW的目标是根据上下文预测中心词的概率,skip-gram模型则相反。一般而言,skip-gram模型的效果会更好,因此本文选择使用skip-gram模型学习站点的向量表示。

skip-gram虽然是一种无监督的方法,但它仍然在内部定义了一个辅助的预测任务。如图5所示,首先选定中心词s_i,在它的前后各m个词距内选定上下文,组成训练的单词对$(s_i, s_{i-m}), \cdots, (s_i, s_{i-1}), (s_i, s_{i+1}), \cdots, (s_i, s_{i+m})$。在训练期间,使用包含当前中心词及其上下文的滑动窗口在语料库中移动,得到所有的训练样本,目的是借助中心词预测上下文出现的概率。构造这个监督学习的目标并不是想要解决这个监督学习问题本身,而是想要借助这一问题来学习一个好的词嵌入模型。

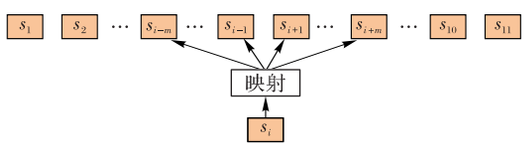


图5 基于skip-gram模型的站点嵌入表示

Fig. 5 skip-gram model based station embedding representation

skip-gram模型的结构如图6所示,是一个简单的神经网络模型,仅拥有输入层、隐藏层和输出层三层结构。

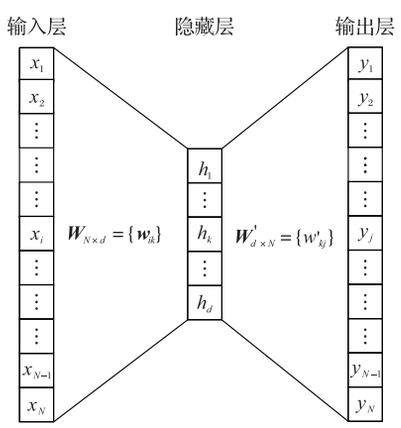


图6 skip-gram模型结构

Fig. 6 skip-gram model structure

输入层和输出层都由一个N维的one-hot编码向量表示,N代表加油站的总数,隐藏层由维度为d的向量表示。权重矩阵$\boldsymbol{W}_{N\times d}$和$\boldsymbol{W}'_{d\times N}$,分别位于输入层与隐藏层之间和隐藏层与输出层之间。隐藏层没有使用任何激活函数,输出层使用softmax作为激活函数。模型的具体的训练过程如下。

1)随机初始化权重矩阵$\boldsymbol{W}$和$\boldsymbol{W}'$;

2)预测目标词的向量$\hat{\boldsymbol{y}}$;

$$\boldsymbol{h} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}$$

$$\boldsymbol{u} = \boldsymbol{W}'^{\mathrm{T}}\boldsymbol{h}$$

$$\hat{y}_j = P\left(s_j \middle| s_i\right) = \frac{\exp\left(u_j\right)}{\sum_{n=1}^{N}\exp\left(u_n\right)}$$

3)通过反向传播算法及随机梯度下降来更新权重矩阵$\boldsymbol{W}$和$\boldsymbol{W}'$,最小化损失函数。整个训练样本集上的损失函数为

$$L = -\sum_{tra\in T}\sum_{s_i\in tra}\left(\sum_{-m\leqslant j\leqslant m, i\neq 0}\log P\left(s_{i+j}\middle| s_i\right)\right);$$

4)取权重矩阵$\boldsymbol{W}$的每一行作为站点的向量表示。

3.3 轨迹聚类

当获得了站点的向量表示之后,就可以利用站点向量得到轨迹的向量表示。在自然语言处理中,句子向量表示的一种简单有效的方法是将句子中所有词向量进行平均[22]。类似地,对于轨迹tra,它的向量表示$\boldsymbol{traj}$为其包含的所有加油站$s_1, s_2, \cdots, s_n$向量表示的平均值,公式表示为:

$$\boldsymbol{traj} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{st}_k$$

至此可以得到每条轨迹的向量表示,然后使用经典的聚类算法进行轨迹聚类。本文采用k均值(k-means)算法进行轨迹聚类。给定样本集D和聚类簇数k,k均值算法将样本集划分为k个不同的聚类簇$C_1, C_2, \cdots, C_k$,最小化平方误差:

$$E = \sum_{i=1}^{k}\sum_{z\in C_i}\left\|\boldsymbol{traj} - \boldsymbol{\mu}_i\right\|_2^2$$

其中:$\boldsymbol{\mu}_i = \frac{1}{|C_i|}\sum_{\boldsymbol{traj}\in C_i}\boldsymbol{traj}$是聚类簇$C_i$的均值向量。$k$均值算法的具体流程如下。

1)从样本集D中随机选择k个样本作为初始聚类簇的均值向量;

2)计算每个样本与各个簇均值向量的距离,选择距离最近的簇,作为样本的簇标记;

3)依据新划分的簇,重新计算簇均值向量;

4)若簇均值向量未发生变化,算法终止,返回聚类簇;否则,重复步骤2)~3)。

4 实验与结果分析

4.1 实验配置

为了挖掘新疆自驾车旅游路线,采用覆盖新疆全区的加油数据集作为原始数据集,该数据集记录了2016年1月1日至2018年12月31日3年时间内,所有人员在新疆的加油记录。首先对原始加油数据集进行预处理(详见第2章),挖掘出20 646名自驾车游客的加油轨迹作为实验数据集,涉及到全新疆1 856个加油站。实验机器系统为Ubuntu 18.04, CPU型号为Intel Core i7-3770 CPU @ 3.4 GHz,内存12 GB, Python版本3.6。

本文的主要参数包括:窗口大小m,站点向量维度d和聚类簇数k。其中,m和d设置为默认值,分别为5和100。新疆

的自驾游路线主要分为北疆线和南北疆大环线,因此将k值设置为2。

4.2　站点分布式表示结果分析

为了验证算法是否学习到了有效的站点嵌入表示,本文文使用余弦相似度衡量站点向量之间的相似性,观察相似的站点之间具有怎样的关联关系。本文选择阿勒泰地区的喀纳斯加油站作为目标,该站点位于著名景区喀纳斯附近,适合分析游客的加油行为。表1展示了与喀纳斯加油站最相似的10个站点。其中的1号、2号、4号、6号和8号站点同样位于喀纳斯景区附近,表明站点的向量表示有效学习到了站点的空间位置特征。除此之外,还发现10个站点均位于著名旅游景点附近,3号、5号和10号加油站位于白沙湖景区附近,7号和9号加油站位于那拉提景区附近。这种结果表明站点向量有效地包含了站点的语义信息,将游客旅行中访问过的景点信息包含在内,有利于进一步的路线挖掘。

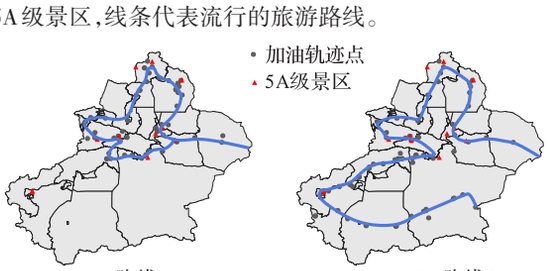
表1　与喀纳斯加油站最相似的10个站点

Tab. 1　Ten most similar stations to Kanas gas station

加油站编号	加油站名称	余弦相似度
1	阿勒泰布尔津东郊加油站	0.936 8
2	阿勒泰北屯西北路加油站	0.847 6
3	阿勒泰哈巴河城西加油站	0.819 1
4	阿勒泰布尔津冲乎尔加油站	0.802 3
5	阿勒泰农十师天山路加油站	0.791 6
6	阿勒泰阿拉哈克加油站	0.788 3
7	伊犁尼勒克乔尔玛加油站	0.787 8
8	阿勒泰布尔津幸福路加油站	0.787 1
9	伊犁新源天河加油站	0.785 0
10	阿勒泰哈巴河齐巴尔加油站	0.781 3

4.3　聚类可视化结果分析

由于数据本身完全无标注,因此选择通过聚类结果的可视化,人工验证结果的有效性。如图7所示,图中的圆点代表每个聚类簇中加油次数前100的站点,三角形代表新疆的12个5A级景区,线条代表流行的旅游路线。



(a) 路线A　　(b) 路线B

图7　聚类结果可视化

Fig. 7　Visualization of clustering results

可以看出,北疆的旅游资源更为丰富和集中,几乎所有游客都会前往北疆进行游览,因此两条路线都涉及到了北疆景点的游览,二者在北疆地区的行程几乎一致,包含了北疆的10个5A级景区。相比北疆,南疆的旅游景点集中在喀什地区,与其他景点相隔较远,因此一部分游客在游览完北疆后选择直接返程(路线A),其余的游客则继续前往喀什地区游览,最终经由巴州的若羌县前往青海省(路线B)。图8为马蜂窝旅游网推荐的新疆自驾游路线,可以看出该路线与图7中的路线A有着很高的重合度,不同之处在于马蜂窝推荐路线以乌鲁木齐市为起止点,这是因为旅游网站通常将乌鲁木齐作为游客的集结地点。

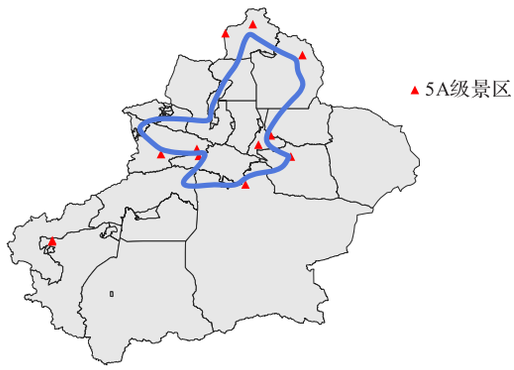


图8　马蜂窝旅游网推荐的一条自驾游路线

Fig. 8　A self-driving tour route recommended by Mafengwo website

5　结语

本文利用用户在新疆的加油数据,成功挖掘出两条流行的新疆自驾车旅游路线。首先,根据自驾车游客的行为特征从全疆加油数据集中挖掘出自驾车游客人群,并分析了游客的基本特征,结果说明了游客群体数据的可靠性。然后,鉴于现有轨迹聚类算法不能解决加油轨迹过于稀疏的问题,提出一种基于语义表示的加油轨迹聚类算法,最终的可视化结果表明该方法很好地还原了游客的旅行路线。但是本文方法仅仅考虑了轨迹点的语义信息,没有将轨迹点本身的空间信息考虑在内,后续的研究工作将考虑把空间与语义信息进行结合,学习更好的站点及轨迹分布式向量表示。另外,本文挖掘出的流行路线为长途自驾游路线,后续的工作将结合节日、季节等信息进一步挖掘短途的自驾游路线。

参考文献(References)

[1] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proceedings of the 2013 Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2013: 3111-3119.

[2] 常亮,孙文平,张伟涛,等. 旅游路线规划研究综述[J]. 智能系统学报, 2019, 14(1): 82-92. (CHANG L, SUN W P, ZHANG W T, et al. Review of tourism route planning[J]. CAAI Transactions on Intelligent Systems, 2019, 14(1): 82-92.)

[3] ZHENG Y, ZHANG L, MA Z, et al. Recommending friends and locations based on individual location history [J]. ACM Transactions on the Web, 2011, 5(1): No. 5.

[4] ZHENG Y, ZHANG L, XIE X, et al. Mining interesting locations and travel sequences from GPS trajectories [C]// Proceedings of the 18th International Conference on World Wide Web. New York: ACM, 2009: 791-800.

[5] ZHENG V W, ZHENG Y, XIE X, et al. Collaborative location and activity recommendations with GPS history data [C]// Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010: 1029-1038.

[6] CUI G, LUO J, WANG X. Personalized travel route recommendation using collaborative filtering based on GPS trajectories[J]. International Journal of Digital Earth, 2018, 11(3): 284-307.

[7] JIANG S, QIAN X, SHEN J, et al. Author topic model-based collaborative filtering for personalized POI recommendations [J].

IEEE Transactions on Multimedia, 2015, 17(6): 907-918.

[8] WEI L Y, ZHENG Y, PENG W C. Constructing popular routes from uncertain trajectories [C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012: 195-203.

[9] 刘艳平,保继刚,黄应淮,等. 基于GPS数据的自驾车游客时空行为研究——以西藏为例[J]. 世界地理研究, 2019, 28(1): 149-160. (LIU Y P, BAO J G, HUANG Y H, et al. Study on spatio-temporal behaviors of self-driving tourists based on GPS data: a case study of Tibet[J]. World Regional Studies, 2019, 28(1): 149-160.)

[10] ZHENG Y. Trajectory data mining: an overview [J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6(3): No. 29.

[11] LEE J G, HAN J, WHANG K Y. Trajectory clustering: a partition-and-group framework [C]// Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2007: 593-604.

[12] TANG W, PI D, HE Y. A density-based clustering algorithm with sampling for travel behavior analysis [C]// Proceedings of the 2016 International Conference on Intelligent Data Engineering and Automated Learning, LNCS 9937. Cham: Springer, 2016: 231-239.

[13] BESSE P C, GUILLOUET B, LOUBES J M, et al. Review and perspective for distance-based clustering of vehicle trajectories [J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(11): 3306-3317.

[14] YI B K, JAGADISH H V, FALOUTSOS C. Efficient retrieval of similar time sequences under time warping [C]// Proceedings of the 14th International Conference on Data Engineering. Piscataway: IEEE, 1998: 201-208.

[15] VLACHOS M, KOLLIOS G, GUNOPOULOS D. Discovering similar multidimensional trajectories [C]// Proceedings of the 18th International Conference on Data Engineering. Piscataway: IEEE, 2002: 673-684.

[16] CHEN L, ÖZSU M T, ORIA V. Robust and fast similarity search for moving object trajectories [C]// Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2005: 491-502.

[17] WANG J, HUANG P, ZHAO H, et al. Billion-scale commodity embedding for e-commerce recommendation in Alibaba [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2018: 839-848.

[18] KENTHAPADI K, LE B, VENKATARAMAN G. Personalized job recommendation system at LinkedIn: practical challenges and lessons learned [C]// Proceedings of the 11th ACM Conference on Recommender Systems. New York: ACM, 2017: 346-347.

[19] RADOSAVLJEVIC V, GRBOVIC M, DJURIC N, et al. Smartphone app categorization for interest targeting in advertising marketplace [C]// Proceedings of the 25th International Conference Companion on World Wide Web. New York: ACM, 2016: 93-94.

[20] GRBOVIC M, CHENG H. Real-time personalization using embeddings for search ranking at Airbnb [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2018: 311-320.

[21] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701-710.

[22] WIETING J, BANSAL M, GIMPEL K, et al. Towards universal paraphrastic sentence embeddings [EB/OL]. [2019-09-20]. https://arxiv.org/pdf/1511.08198.pdf.

This work is partially supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2019D01A92), the Tianshan Distinguished Young Scholars of Xinjiang (2018Q005).

YANG Fengyi, born in 1994, M. S. candidate. His research interests include big data analysis, data mining.

MA Yupeng, born in 1979, Ph. D., research fellow. His research interests include Internet of things, big data analysis.

BAO Hengbin, born in 1995, M. S. candidate. His research interests include big data analysis, data mining.

HAN Yunfei, born in 1990, Ph. D., assistant research fellow. His research interests include data mining, computer vision.

MA Bo, born in 1984, Ph. D., associate research fellow. His research interests include big data analysis, knowledge graph.