ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 3,May 2018,pp. 538－546
DOI:10. 16337/j. 1004-9037. 2018. 03. 018

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种基于位置语义和概率的人群分类方法

邱运芬[1]　张晖[1]　李波[1,2]　杨春明[1]　赵旭剑[1]

(1.西南科技大学计算机科学与技术学院,绵阳, 621010;2.中国科学技术大学计算机科学与技术学院,合肥, 230027)

摘　要:针对现有的人群分类方法忽略地理位置隐含的功能特征及其访问概率的问题,提出了一种基于位置语义和概率的人群分类方法。该方法主要包括位置语义发现和访问概率向量聚类两部分:首先,采用位置语义发现方法得到位置词汇所隐含的位置语义;其次根据位置语义分配情况获得移动用户对位置语义空间的访问概率向量;最后将其作为聚类分析的权向量,实现人群分类。实验结果表明,该方法提取出的位置语义与现实相符,得到的同类用户在位置语义空间的访问概率向量相似。与现有的人群分类方法相比,本文提出的人群分类方法 F-measure 值提高了 4%,实验效果更优。

关键词:人群分类;位置;语义信息;概率向量;聚类分析

中图分类号:TP391　　**文献标志码**:A

Group Classification Method Based on Location Semantic and Probability

Qiu Yunfen[1], Zhang Hui[1], Li Bo[1,2], Yang Chunming[1], Zhao Xujian[1]

(1. School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, 621010, China;
2. School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230027, China)

Abstract: The existing group classification methods ignore the functional characteristics and their access probabilities implied in geographical positions. To solve this problem, a group classification method based on location semantic and probability is proposed, which includes two parts: the location semantic discovery and the access probability vector clustering. Firstly, the location semantic implied in location words is obtained by using location semantic discovery method. Then according to the location semantic distribution, the access probability vector of mobile users for the location semantic space can be obtained. Finally, the group classification can be realized by using the access probability vector as the clustering weight vector. Experimental results show that the proposed method can effectively extract the location semantic coinciding with the reality and obtain similar users with similar access probabilities in location semantic space. Compared with the available group classification methods, the proposed method can achieve better experimental effects with an increase in F-measure of 4%.

Key words: group classification; location; semantic information; probability vector; cluster analysis

---

基金项目:赛尔网络下一代互联网技术(NGII20150510)资助项目。
收稿日期:2016-09-06;**修订日期**:2016-10-28

引 言

现代城市由各种各样的功能区域组成,人们每天在这些功能区域中进行不同的社会活动,如购物、上下班、生活和旅游等。同时,随着定位服务的兴起,基于位置服务的应用软件迅猛发展,可以通过位置服务 APP 获取民众在这些功能区域中活动产生的 GPS 坐标,如社交软件微信、导航软件高德地图等。深入挖掘移动用户 GPS 坐标的功能特征,计算其在该功能特征区域出现的概率大小,可以研究移动用户的兴趣爱好,为判断用户类型奠定坚实的基础。例如,若用户 A 经常出现的 GPS 坐标的功能特征为餐厅,可猜测 A 用户对于饮食文化有一定程度的研究,对于研究用户的兴趣爱好具有重大意义。目前,基于 GPS 坐标的人群分类研究大部分基于出现在相同或地理位置相近的用户通常为同类用户的假设。这种方法具有一定局限性,得到的同类用户基本上都在相同或相近的区域内活动。若用户 B 经常出入地方的 GPS 坐标与用户 A 经常出入的 GPS 坐标距离相差较大,该方法会认为 A 与 B 不是同类用户。但考虑现实情况,若 A 与 B 产生的 GPS 坐标具有相同的功能特征(如餐厅),他们即为同类用户。故部分学者提出了基于功能特征的人群分类方法,但这部分研究仅局限于探究用户是否拥有相同的功能特征,却忽略了用户访问不同功能特征的不确定性。针对以上问题,本文从功能特征和访问功能特征的不确定性两个方面考虑,从具象 GPS 坐标引申出抽象位置语义的概念,以更高维度解析用户访问 GPS 坐标的目的性与意义,并计算用户访问不同位置语义的概率大小,将用户对不同位置语义的访问倾向作为特征进行人群分类,最终得到人群分类结果。

1 相关工作

随着定位技术的高速发展,基于位置服务的应用软件越来越多,更容易获取用户位置数据,因此越来越多的学者投身到基于位置数据的人群分类研究中。到目前为止,按照人群分类特征选取的不同,可分为两大类:基于 GPS 坐标和基于功能特征。前者认为 GPS 坐标作为移动用户最重要的特征,是辅助人群分类的重要属性,频繁出现在相同或相近地理位置的用户可视为同类人群,因此部分学者采用频繁模式[1-4]挖掘用户频繁出现的位置坐标法,将其作为用户分类特征。宋衡等[5]采用主成分分析法(Principal component analysis,PCA)提取不同用户经常出现的位置坐标,将其作为分类特征,首先收集 3 个年级在校学生的位置数据集,利用 PCA 抽取用户特征,从而对学生进行年级分类。在此基础上,张成等[6]提出了一种基于 PCA 的单变量贡献度方法,其核心思想为利用最大似然估计算法提取用户分类特征,从而对人群进行分类管理。但是如前文所述,基于 GPS 坐标的人群分类算法局限于具象的 GPS 坐标的地域相近性,忽略了用户访问该地理位置的潜在意义。

近年来,很多研究者致力于研究用户访问地理位置的潜在意义,即地理位置隐含功能特征的提取,如发现地区功能特征(Discovering regions of different functions,DRoF)的框架,用于提取地理位置隐含的功能特征[7]。输入移动用户产生的位置数据和先验兴趣点,框架由此计算出移动用户地理位置的功能特征,但该方法的缺陷是需要提前收集用户兴趣点的先验知识,会耗费一定的人力物力。Yuan 等[8]先按城市主干道(如高速公路)对地理位置进行区域划分,然后按照时间顺序连接 GPS 坐标点为用户移动轨迹,在此基础上挖掘功能特征。该方法区域划分粒度不好掌握,如按照高速公路划分区域,会导致功能区域的范围较大。文献[9]从用户的行为出发,认为用户行为与功能特征密切相关,利用移动用户在该区域内的手机行为(电话、各类 APP 使用情况等)推断功能特征。

随着功能特征提取方法的逐渐完善,基于功能特征的各类研究也逐渐成为热门,其中基于功能特征的人群分类也得到部分学者的青睐。Lee 等[10]提出了多项传播率(Multiple propagation rate,MPR)算法,该方法抽取用户频繁出现的 K 个地理位置,并利用用户手机 APP 的使用情况构建地理位置分类层次图,从中获取 K 个地理位置代表的功能特征,将这两者作为用户特征进行用户相似性计算。该方法

只选取用户的 K 个频繁点，忽略了总体功能特征访问次数大的地理位置，会造成一定的误差，其次手动构建地理位置分类层次图工作量较大。Xiao[11,12] 考虑用户在功能特征间移动的先后顺序，抽取用户移动轨迹，采用最大序列算法计算移动用户移动轨迹的相似度，但在构建用户移动轨迹时并没有考虑用户在不同功能特征地区的出现概率，因此构建的移动轨迹含有不能体现用户生活习惯的点。

针对人群分类的现有问题，本文提出了一种基于位置语义和概率的人群分类方法：首先利用贝叶斯思想，位置语义的分布满足多项分布，迭代求出位置词汇下隐藏的位置语义分布；然后在得到位置语义分配的情况下，选出权重最高的前 20 个位置词汇，借助百度地图查看位置语义指代的具体含义，如生活区等；最后将用户在位置语义空间下的访问概率向量作为聚类特征向量，找到同类用户，并根据位置语义指代的具体含义确定人群类型。

2 人群分类方法

定义 1 位置词汇（Location word，LW），用户的 GPS 坐标，由经纬度唯一标示，具有唯一性，表示为 $p=\{p_L, p_R\}$（p_L 表示经度，p_R 表示纬度）。

定义 2 位置语义（Location semantic，LS），位置词汇指代的功能特征，表示为 z，$z\in\{z_1, z_2, \cdots, z_T\}$。其中，$z_i$ 表示具体的位置语义，T 为位置语义总数。

基于以上两个定义，本文提出的人群分类方法主要分为两部分：位置语义发现和访问概率向量聚类，图 1 给出了其流程图。如图 1(a)所示，输入 4 个用户的位置数据集，输出位置词汇指代的位置语义。User 1 和 User 2 虽然分别出现在不同的地理位置（茗缘茶楼和尚雅咖啡），但却同时拥有为餐饮区的位置语义；同理，User 3 和 User 4 同时拥有为住宅区的位置语义。经过第 1 步后，计算用户在位置语义空间的访问概率向量（见图(b)）。从图 1 可知，User 1 和 User 2 对住宅区和教学区这两种位置语义的访问概率偏大，而 User 3 和 User 4 对餐饮区这一位置语义的访问概率偏大。最后将访问概率向量作为聚类特征计算用户相似度，得到同类用户。

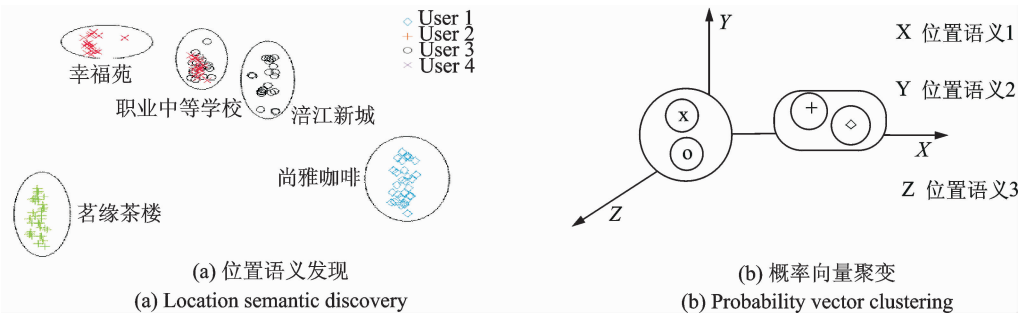

图 1 算法流程图

Fig. 1 Algorithm flowchart

2.1 位置语义发现

假设输入的位置集合中包含 T 个位置语义，从用户位置文档 m 中随机抽取一个位置词汇 p，其位置语义表示为 z_i，$i\in\{1,2,\cdots,T\}$，重复 N 次后，位置语义 z_i 的抽取次数表示为 n_i，则位置语义抽取结果 $z=(n_1, n_2, \cdots, n_T)$ 满足多项分布，同理，$Pr(z)\equiv\left(\frac{n_1}{N}, \frac{n_2}{N}, \cdots, \frac{n_T}{N}\right)$ 满足多项分布。已知多项分布与狄利克雷分布具有共轭性，因此，可将狄利克雷分布作为位置语义抽取的先验分布，并表示为 λ。

位置语义发现主要分为 3 步：

(1)位置语义抽取满足多项分布，由 λ 的概率密度可知用户在位置语义空间出现的概率 $\boldsymbol{\theta}_m$ 满足

$$Pr(\boldsymbol{\theta}_m|\lambda)=\frac{\Gamma\left(\sum_{k=1}^{T}\lambda_k\right)}{\prod_{k=1}^{T}\Gamma(\lambda_k)}\prod_{k=1}^{T}\boldsymbol{\theta}_{m,k}^{\lambda_k-1} \tag{1}$$

式中:$\Gamma(x)$为 gamma 函数。

(2)设定变量 ε,基于 ε 和 $\boldsymbol{\theta}_m$ 的取值,提取每个位置词汇的位置语义。ε 的取值满足

$$\varepsilon=\begin{cases}1 & \mathrm{dis}(p,p_{\mathrm{targ}})\geqslant \mathrm{distance}\\ 0 & \mathrm{dis}(p,p_{\mathrm{targ}})<\mathrm{distance}\end{cases} \tag{2}$$

式中:p_{targ}表示目标位置词汇,p 表示当前位置词汇,$\mathrm{dis}(p,p_{\mathrm{targ}})$表示 p 与 p_{targ}的物理距离,distance 表示距离阈值。依次遍历每个用户文档,ε 初始值为 1。当 p_{targ}与 p 的物理距离大于 distance 时,设置 $\varepsilon=1$,同时利用 $\boldsymbol{\theta}_m$ 为 p 重新分配一个位置语义,并设置 $p_{\mathrm{targ}}=p$;反之,则设置 $\varepsilon=0$,同时认为 p_{targ}与 p 具有相同的位置语义。这样操作的意义是,既能保证相近的位置词汇必定属于同一位置语义,又能让距离较远的位置词汇有机率获得相同的位置语义,符合现实情境。按照经验,用户在某个位置语义内的活动范围一般较集中,同时为避免距离阈值设值太大造成误差过大,本实验中的 distance 取值为 50。

当 $\varepsilon=1$ 时,需计算 $Pr(z|p)$,为位置词汇 p 取其概率最大的位置语义重新分配。由全概率公式和贝叶斯公式[13]可知:$P_r(z|p)=\frac{Pr(z,p)}{\sum_z Pr(z,p)}$。由于分母 $Pr(z|p)$的求解非常困难,因此采用计算较为简单的 Gibbs 采样算法[14,15]求得 $Pr(z|z_{\neg i},p)$,以此来近似 $Pr(z|p)$。$Pr(z|z_{\neg i},p)$的计算公式可表示为

$$Pr(z=k|z_{\neg i},p)=\frac{Pr(z,p)}{Pr(z_{\neg i},p)}=\frac{n_{k,\neg i}^{(t)}+\beta}{\sum_{t=1}^{V}n_{k,\neg i}^{(t)}+V\beta}\cdot\frac{n_{m,\neg i}^{(k)}+\lambda}{\sum_{k=1}^{T}n_{m,\neg i}^{(k)}+T\lambda} \tag{3}$$

式中:$n_{m,\neg i}^{(k)}$表示位置文档 m 下位置语义 k 出现的次数($k\neq z_i$);$n_{k,\neg i}^{(t)}$表示位置语义 k 下位置词汇 t 出现的次数($t\neq i$);β 表示位置词汇的狄利克雷先验分布。

(3)由式(3)为位置词汇执行分配操作,统计位置词汇在位置语义下的出现次数,使用狄利克雷期望公式[16]更新 $\boldsymbol{\theta}_m$,并重复此步骤,以达到用户访问概率向量收敛,即

$$\boldsymbol{\theta}_{m,k}=\frac{n_m^{(k)}+\lambda}{\sum_{k=1}^{T}n_m^{(k)}+T\lambda} \tag{4}$$

以此得到每个位置词汇的位置语义和每个用户在位置语义空间下的访问概率向量。

2.2 访问概率向量聚类

由定义 2 可知,位置语义暗示着用户出现在该区域的目的性,表示位置词汇隐含的功能特征,同时,访问概率向量表示移动用户在位置语义空间的出现概率,暗含用户在该区域出现的不确定性。因此,将位置语义和访问概率向量共同作为用户相似性计算标准,既考虑了用户出现在地理位置的深层含义,不再局限于坐标位置的地理限制,也考虑了用户访问不同位置语义的不确定性。如某用户包含一系列位置语义 $\boldsymbol{z}=\{z_1,z_2,z_3,z_4\}$,分别代表教学区、住宅区、餐饮区和娱乐区,与该位置语义空间对应的访问概率向量为 $\boldsymbol{\theta}=\{0.4,0.5,0.05,0.05\}$。综合两者来看,该用户在日常生活中,有访问过 4 种类型的位置语义,但在教学区和住宅区出现的概率最大,在餐饮区和娱乐区访问概率较小,从而可作出较为合理的推测:该用户可能为学生或教职员工。因此,若要寻找该用户的同类用户,也应包含相同的位置语义,且具有相似的访问概率向量。因此,将用户 m 在特定位置语义空间的访问概率向量作为人群聚类的特征向量,即有

$$\boldsymbol{\theta}_m=\{Pr(z_1),Pr(z_2),\cdots,Pr(z_T)\} \tag{5}$$

式中:$Pr(z_i)$表示用户在位置语义 z_i 出现的概率,且 $Pr(z_1)+Pr(z_2)+\cdots+Pr(z_T)=1$。使用通用聚

类算法对访问概率向量聚类得到的结果即为人群分类结果。

3 位置语义发现和人群分类实验

3.1 实验数据及数据预处理

本文历时两个月(2015-08-13 至 2015-10-10),收集了某地区的移动用户通过使用位置服务类 App 所产生的位置数据。收集的数据属性包括经度、纬度和 App 名称等信息,其中经度和纬度共同组成位置词汇,APP 用于后期用户类型标识,具体说明如表 1 所示。

表 1 数据格式说明

Tab. 1 Data format description

Longitude	Latitude	Object ID	APP Name
地理位置经度	地理位置纬度	用户 ID	APP 名称

在进行实验之前,需要先对数据进行预处理,避免误差数据影响实验结果。数据筛选包括两类:(1)异常点去除:只保留某地区范围内的位置词汇,过滤掉其他范围的位置词汇,以免造成数据混淆和增大位置语义标识的难度;(2)数据选取:随机抽取 1 000 个用户,以保证数据选取的随机性。约 33 万条位置记录进行后期实验,保证数据量充足和实验结果的准确性。

3.2 评价指标

现有的人群分类方法,用户类型的判断大多采用人工标注,在准备实验数据时需耗费大量精力,且受人为因素影响较大。因此,本文从两个方面对实验结果进行评价:内部评价和外部评价。内部评价用于评估访问概率向量在各类聚类方法中的聚合度[17],外部评价则用于评估人群分类结果的正确性,两个评价方式的计算公式分别如式(6,7)所示。

(1) 内部评价指标:Dunn index

$$D=\frac{\min\limits_{1\leqslant i\leqslant j\leqslant n} d(i,j)}{\max\limits_{1\leqslant k\leqslant n} d'(k)} \tag{6}$$

式中:分母表示取分类 k 中任意两个移动用户的相似度 $d'(k)$ 的最大值;分子表示取类别 i 和 j 的相似度 $d(i,j)$ 的最小值。D 值越高,意味着同簇内用户相似度越高,簇间用户相似度越低,即达到最佳聚合结果,因此 D 值越高表示聚合度越高。但 D 值的大小并不能判定人群分类结果的正确性,因此,本文引入了外部评价指标用于评价人群分类正确性。

(2) 外部评价指标:APP 类标签。通过对位置数据集和相关研究的深入分析,可知位置语义与产生位置词汇的 APP 之间存在着一定的关联关系[9]。用户处在不同的位置语义下,会有不同的手机行为,比如,如果用户处于餐饮区,用户则可能使用美团 APP,便于搜索附近美食或参与团购。基于此认识,将采集到的位置数据中的 APP 名称属性作为标注用户类型的依据。采用 F-measure 指标评价人群分类结果的优劣,其计算公式为

$$F=\frac{2*P*R}{P+R} \tag{7}$$

式中 P 和 R 分别表示准确率和召回率。

位置数据集中共包含 21 种常用 APP 名称,将其分成 5 大类,具体如表 2 所示。由于 APP 名称出现的次数可近似表示用户访问位置语义的概率大小,因此可根据每个用户位置文档中每种类型 APP 名称出现的次数来决定用户所属类型,其计算公式为

$$C=\max\{N_{\mathrm{t}},N_{\mathrm{f}},N_{\mathrm{e}},N_{\mathrm{j}},N_{\mathrm{c}}\} \tag{8}$$

表 2 APP 类标签

Tab. 2 APP class labels

类标签	APP 名称
商业型	手机淘宝 58 同城 京东商城 国美在线 我查查 旺信
餐饮娱乐型	大众点评 百度糯米 美团
旅游型	墨迹天气 淘宝酒店 百度旅游 阿里旅行
居家教育型	搜狐新闻 今日头条 新浪微博 暴风影音 腾讯 芒果 TV
运输型	百度地图 老虎地图

式中:N_t,N_f,N_j,N_e 和 N_c 分别表示表 2 中 5 种类型 APP 名称出现的次数,取其最大值作为用户类型标签。

3.3 语义数目选择

在位置语义的提取过程中,语义数 T 的选择对实验结果及性能影响甚大,因此需要通过实验预先确定其大小。采用困惑度[18]来确定 T 值,其计算公式为

$$\text{Perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{w=1}^{W}\log(Pr(p_w))}{\sum_{w=1}^{W}N_w}\right\} \tag{9}$$

式中:W 表示测试集文档数目,N_w 表示测试位置文档 w 的位置词汇总数;分母表示 W 个文档的位置词汇总数;$P_r(P_\omega)$表示 p_w 的产生概率。实验中,先验分布 $\lambda=50/T$ 为初始值,并且将语义数 T 分别设置每次新增 5,采用 Gibbs 采样[19],分 8 次实验分别计算困惑度,取其合适的语义数作为后续实验前提,实验结果如图 2 所示。

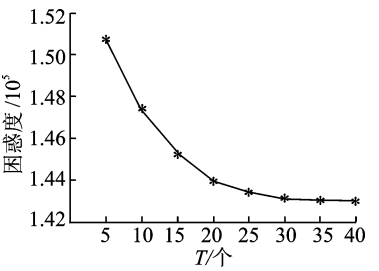

图 2 位置语义发现方法的困惑度

Fig. 2 Perplexity of location semantics discovery method

从图 2 可以看出,困惑度随着语义数 T 的增大而逐渐降低,最后在[25,40]区间趋于稳定。当困惑度越低时,表示模型的泛化能力越强,但同时位置语义数目作为访问概率向量的维数,不宜取值过大,维数过大会影响计算效率。综上两点,位置语义数目取值为 30 较为合适。

3.4 实验结果与分析

3.4.1 位置语义实验结果

从每个位置语义下选择 20 个权重最大的位置词汇,借助百度地图,查看每个位置词汇的具体位置语义,得到位置语义的具体含义。表 3 展示了其中 6 个位置语义和其权重排名前 5 的位置词汇。当位置词汇在百度地图中展示为住宅区时,位置语义可解释为住宅区;当位置词汇在百度地图中展示为娱乐休闲场时,位置语义可解释为娱乐区,其他类型位置语义以此类推。从实验结果看出,不同的位置词汇可能拥有相同的位置语义,且本文提出的位置语义发现方法得到的实验结果能准确表达位置词汇所具有的功能特征。

3.4.2 人群分类结果

选择基于位置词汇的人群分类算法 PCA[5],基于功能特征的人群分类算法 MPR[10]与本文提出的方法进行对比实验。在 MPR 算法中,参考原文,取频繁出现的 50 个位置词汇及其位置语义作为分类特征。

为了更全面地比较 3 种特征选取方法的优劣,选取 4 种普遍的聚类算法,包括划分聚类 K 均值(K-means)、密度聚类(Density-based spatial clustering of applications with noise,DBSCAN)、层次聚类(Hierarchical clustering,HC)和吸引力传播聚类(Affinity propagation,AP)[20],尽可能忽略因聚类算法造成的误差,对比 3 种方法得到的人群分类结果。如 3.2 节所述,分别采用 Dunn index 和 F-measure 作为内

部和外部评价指标，4 种聚类方法得到的 Dunn index 值和 F-measure 值分别如图 3 和图 4 所示。

表 3 位置语义下的位置词汇

Tab. 3 Location words of location semantic

位置语义	GPS 坐标
住宅区	刘家湾小区(105.545 954,30.540 4) 滨江小区(105.397 328,30.879 252) 新塘房小区(105.614 108,30.528 601) 奥城花园(105.596 632,30.532 626) 金科美湖湾(105.621 354,30.511 078)
教学区	实验学校(105.584 135,30.500 908) 射洪旅游学校(105.384 035,30.864 254) 长乐街小学(105.592 682,30.511 131) 横山镇小学(105.458 449,30.505 567) 东禅镇初中(105.344 655,30.329 781)
度假区	东方生态旅游度假区(105.250 6,30.578 71) 东方生态旅游度假区(105.248 908,30.580 221) 死海旅游度假区(105.254 676,30.602 39) 龙凤园林(105.354 909,30.909 242) 西山风景带(105.589859,30.493412)
娱乐区	天娱音乐会所(105.569 315,30.527 051) 立品酒家(105.597 412,30.497 398) 九九休闲山庄(105.577 36,30.506 522) 平安寨(105.385 47,30.897 68) 和平山庄(105.573 83.30.511 2)
商业区	三清街(105.585,30.512 155) 以纯(105.718 49,30.781 048) 明月路(105.574 908,30.534 459) 昆仑好客(105.539 226,30.538 831) 圣帝保罗(105.564 995,30.471 746)
医院	安康康复医院(105.458 529,30.348 503) 蓬溪县人民医院(105.714 469,30.779 513) 爱心药业(105.557 890,30.544 867) 安居区人民医院(105.458 963,30.346 019) 安居区人民医院(105.457 131,30.345 894)

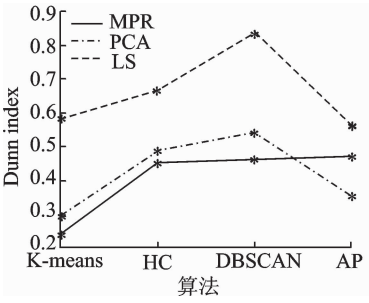

图 3 4 种算法的 Dunn index 对比

Fig. 3 Dunn indexes comparison of four algorithms

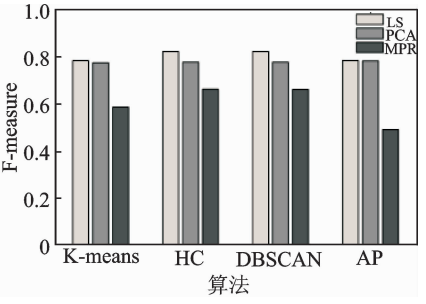

图 4 4 种算法的 F-measure

Fig. 4 F-measure of four algorithms

由图 3 可看出，本文提出用于人群分类的特征聚类后得到的聚合度最高，说明本文提出的访问概率向量更利于分类特征的聚合。但如上所述，Dunn index 值只能说明本文提出的分类特征更利于聚合，却不能对人群分类结果进行评判。因此，需要用 F-measure 从另一个方面来评价人群分类结果的优劣。由图 4 可看出，本文方法得到的 F-measure 高于另外两种方法。

PCA 方法仅能得到访问区域相近的同类人群，对拥有相同位置语义和不同位置词汇的用户不能判断为同类人群，因此分类效果不尽如人意。而 MPR 算法只抽取用户频繁出现的前 50 个位置词汇，得到其位置语义，综合考虑位置词汇和位置语义，将其作为用户特征计算相似性，但在本文中位置词汇是 GPS 坐标，粒度很小，且并没有作精度处理，因此用户频繁访问的 GPS 坐标完全相同的可能性很低。因此用户频繁出现的前 50 个位置词汇出现的次数都偏低，并不能完全体现出用户对该位置词汇的频繁访问。而本文的人群分类方法同时考虑位置语义和访问概率两方面，将具象的位置词汇抽象成更高维度的位置语义，挖掘位置词汇更深层的含义，加入用户访问位置词汇的意图，不再依赖于判断细粒度 GPS 坐标的相似性，提高了人群分类结果的召回率；同时将用户对位置语义空间的访问概率向量作为聚类特

征,不再仅依赖于判断用户是否访问过相同的位置语义,并引入用户对位置语义访问概率的不确定性,从而提高了人群分类的准确率。

3.4.3 人群分类结果解释

图 5 和图 6 采用更直观的方式展示了基于位置语义与位置词汇进行人群分类的明显区别。由内部评价 Dunn index 可看出,DBSCAN 的聚合度最高,所以选取 LS-DBSCAN 和 PCA-DBSCAN 作为对比。图 5 为采用 LS-DBSCAN 得到的同类用户,图 6 为采用 PCA-DBSCAN 得到的同类用户。从图中可看出,User 2 和 User 3 属于物理意义上的相似用户,访问的位置词汇大多距离相近或相同;而 User 1 和 User 2 属于位置语义和访问概率相近的同类用户,更符合现实意义。由此可知,本文的人群分类方法具有更高的召回率和准确率。

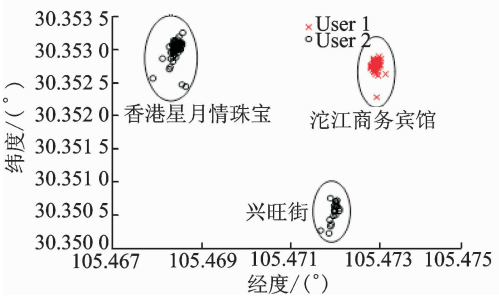

图 5 LS-商业型用户

Fig. 5 LS-commercial users

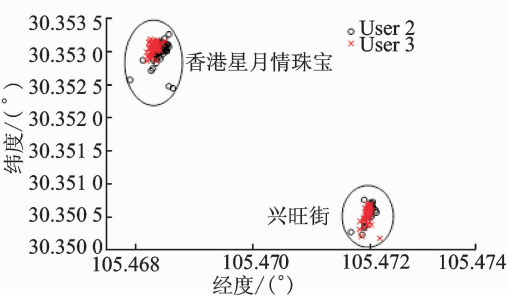

图 6 PCA-商业型用户

Fig. 6 PCA-commercial users

4 结束语

人们在各个功能区域中活动产生的 GPS 坐标是用户类型判断的重要依据,且用户在不同功能区域,会有不同的手机操作行为。深入挖掘移动用户 GPS 坐标的位置语义,研究用户访问不同位置语义的概率倾向,对于研究群体用户的兴趣爱好和用户类型具有重要意义。基于 GPS 坐标的人群分类方法按用户活动区域进行人群划分,得到的同类用户都出入在相同或相近的位置区域,为物理意义上的相似用户;而现有基于功能特征的人群分类局限于判断用户是否拥有相同的位置语义,忽略了用户对位置语义访问的不确定性,没有全面考虑用户在位置语义空间的出现概率。针对上述问题,本文提出了一种基于位置语义和概率的人群分类方法。该方法首先通过位置语义发现方法挖掘位置语义,实验结果表明,该方法得到的位置语义能较准确地说明位置词汇的功能特征;然后结合位置词汇的位置语义分配情况,计算用户在位置语义空间上的访问概率向量,考虑用户在不同位置语义上的访问倾向;其次将用户的访问概率向量作为聚类矩阵,采用聚类方法计算用户间相似度得到同类用户;最后根据位置语义的具体含义,标识用户类型。将位置语义与访问概率向量结合作为人群分类的特征与现有的方法相比具有更高的 F-measure 值。今后的研究工作将把时间属性加入到位置语义中,抽取位置语义随时间的变化轨迹,进一步挖掘用户在时间维度上的行为模式,并比较用户行为模式间的相似性。

参考文献:

[1] Xue Andyyuan, Zhang Rui, Zheng Yu, et al. Destination prediction sub-trajectory synthesis and privacy protection against such prediction [C] //Proceedings of the 29th IEEE International Conference on Data Engineering. Brisbane: IEEE, 2013: 254-265.

[2] Zheng Kai, Zheng Yu, Yuan N J. Discovery of gathering patterns from trajectories [C] //Proceedings of the 29th International Conference on Data Engineering. Brisbane: IEEE, 2013:242-253.

[3] Tang Lu-an, Zheng Yu, Yuan Jing, et al. On discovery of traveling companions from streaming trajectories [C] //Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. Washington: IEEE, 2012:186-197.

[4] Sheng Chang, Zheng Yu, Hsu Wynne, et al. Answering top-k similar region queries [C] //Proceedings of the 15th International Conference on Database Systems for Advanced Applications. Japan: Springer, 2010:186-201.

[5] 宋衡. 基于位置数据的人类行为识别和相似性研究[D]. 上海: 上海交通大学, 2014.
Song Heng. Human behavior recognition and similarity analysis based on location data[D]. Shanghai: Shanghai Jiaotong University, 2014.

[6] 张成, 刘亚东, 谢彦红, 等. 基于 PCA 与 MLE 方法的人群分类新方法研究[J]. 沈阳化工大学学报(自然科学版), 2015, 29(2):168-171.
Zhang Cheng, Liu Yadong, Xie Yanhong, et al. A new method of population classification based on PCA and MLE[J]. Journal of Shengyang University of Chemical Technology(Natural Science Edition), 2015, 29(2):168-171.

[7] Yuan Jing, Zheng Yu, Xie Xing. Discovering regions of different functions in a city using human mobility and POIs [C] //Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012:186-194.

[8] Yuan Nicholas Jing, Zheng Yu, Xie Xing, et al. Discovering urban functional zones using latent activity trajectories [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3):712-725.

[9] Toole J L, Ulm M, Gonzalez M C, et al. Inferring land use from mobile phone activity [C] //Proceedings of the ACM SIGKDD International Workshop on Urban Computing. New York: ACM, 2012:1-8.

[10] Lee M J, Chung C W. A user similarity calculation based on the location for social network services [C] //Proceeding of the 16th International Conference on Database Systems Advanced Applications. Hong Kong: Springer, 2011, 4(1):38-52.

[11] Xiao Xiang Ye, Zheng Yu, Luo Qiong, et al. Finding similar users using category-based location history [C] //Proceedings of the 18th SIGSPATIAL International Symposium on Advances in Geographic Information Systems. New York: ACM, 2010:442-445.

[12] Xiao Xiangye, Zheng Yu, Luo Qiong, et al. Inferring social ties between users with human location history [J]. Journal of Ambient Intelligence and Humanized Computing, 2012, 5(1):3-19.

[13] 蒋铭初, 潘志松, 尤俊. 基于 PLSA 主题模型的多标记文本分类 [J]. 数据采集与处理, 2016, 31(3):541-547.
Jiang Mingchu, Pan Zhisong, You Jun. Multi-label text categorization algorithm based on topic model PLSA[J]. Journal of Data Acquisition and Processing, 2016, 31(3):541-547.

[14] Li Chengtao, Zhang Jianwen, Sun Jiantao, et al. Sentiment topic model with decomposed prior [C] // Proceedings of the 2013 SIAM International Conference on Data Mining. Austin: SIAM, 2013: 767-776.

[15] Lin Chenghua, He Yulan, Richard Everson, et al. Weakly supervised joint sentiment-topic detection from text [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(6):1134-1145.

[16] Chen Zhiyuan, Liu Bing. Mining topics in documents: Standing on the shoulders of big data [C] //Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 1116-1125.

[17] Xu Dongkuan, Tian Yingjie. A comprehensive survey of clustering algorithms [J]. Annals of Data Science, 2015, 2(2):165-193.

[18] Yang Guangbing, Wen Dunwei, Kinshuk, et al. A novel contextual topic model for multi-document summarization[J]. Expert Systems with Applications, 2015, 42(3):1340-1352.

[19] 张俊鹏, 贺建峰. 基于 LDA 主题模型的功能性 miRNA-mRNA 调控模块识别[J]. 数据采集与处理, 2015, 30(1):155-163.
Zhang Junpeng, He Jianfeng. Identifying of funtional minRNA-mRNA regulator modules based on LDA topic model[J]. Journal of Data Acquisition and Processing, 2015, 30(1):155-163.

[20] Brendan J F, Delbert D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814):972-976.

作者简介:

邱运芬(1992-),女,硕士研究生,研究方向:文本挖掘,E-mail: 531802979 @ qq.com。

张晖(1972-),男,博士,教授,研究方向:文本挖掘、知识工程,E-mail: zhanghui@swust edu.cn。

李波(1977-),男,讲师,研究方向:信息过渡、信息安全。

杨春明(1980-),男,副教授,研究方向:文本挖掘、知识工程。

赵旭剑(1984-),男,博士,讲师,研究方向:中文信息处理、Web 信息检索。

(责任编辑:王静)