文章编号:1671 - 6833(2017)02 - 0017 - 04

基于 K-Means 和时间匹配的位置预测模型

胡 燕, 朱晓瑛, 马 刚

(北京邮电大学 信息网络中心,北京 100876)

摘 要: 随着移动服务的发展,越来越多的移动端服务基于对象的位置进行推送和推荐,因此位置预测技术显得越来越重要. 由于对象位置信息存在采集不连续或对象行为不规律等因素,导致位置预测成为一项非常有挑战的工作. 为了提高位置预测的准确性,提出一种基于 K-Means 算法和时间匹配的位置预测模型. 该模型使用 K-Means 算法对历史位置点进行聚类,划分多个对象运动区域,针对对象运动区域进行预测. 按照对象的作息时间将一天时间划分为多个时间段,运用笔者提出的轨迹建模算法和轨迹更新算法形成用户运动轨迹,形成对象运动轨迹,再使用时间匹配原则进行位置预测. 笔者最后利用真实的数据实现该模型,实验证明:未使用该模型的位置预测准确率为 39.7%;使用该模型后算法和时间匹配的位置预测模型预测准确率达到 60.3%,准确率提高了 20% 左右.

关键词: 位置预测;K-Means 算法;时间匹配;聚类

中图分类号: TP311 **文献标志码:** A doi:10.13705/j.issn.1671 - 6833.2017.02.005

0 引言

随着移动对象空间技术的快速发展,对象位置信息采集更加便捷,随之而来的移动位置服务也越来越重要. 为了使服务具有前瞻性,不仅要对移动对象的位置进行分析,更要对其位置进行预测[1-2]. 对象位置信息可以通过多种方式采集,例如 GPS、WiFi、AP 等,这些位置信息和访问时间对应就能形成一个对象的轨迹信息. 然而,位置预测却是一项有难度的工作,例如位置信息采集总是不连续的或者存在盲点,一些重要位置信息无法采集到;已采集的数据通常无法直接进行对象位置预测,需要预处理工作;对象的行为存在不规律和不确定性,很难用一种算法进行预测,要综合考虑多种因素,以上这些原因导致位置预测比较困难.

位置预测通常有两种方法,第一种是根据对象的上一个访问位置点预测当前位置点,通过计算转移概率进行预测. 文献[3 - 5]中的马尔科夫和隐马尔科夫算法用于位置预测,同时结合对象社会关系和时间匹配,这种预测方式只与上一个位置到当前位置的转移概率有关. 其中,文献[3]将贝叶斯网络用于预测,并且考虑多种因素来提高预测效率,将空间、时间、对象相似性等属性考虑到预测模型中. 文献[6 - 8]使用漫步算法和马尔科夫算法同时进行预测,对象访问路径和时间间隔也作为预测的影响因素. 第二种方式是收集历史位置点信息预测当前位置. 文献[1]和文献[9]对历史活动位置进行建模,并且将对象的运动趋势作为位置预测的重要因素.

笔者的位置预测属于第二种预测方式,即采用基于对象历史访问位置信息进行预测. 由于预测对象的精确位置较为困难,且移动服务推送总是面向某个区域的对象提供服务,因此笔者的预测模型也以区域为基础. 首先对对象的历史位置点用 K-Means 算法进行聚类,每一个聚类的结果就是一个对象活动区域,再基于以上区域的划分进行位置预测以提高预测的准确性,并且对对象所在位置区域的预测对移动服务提供者是有价值的. 最后基于时间匹配原则,将对象 24 h 的时间进行分段,并将每个时间段历史上出现概率最多的位置区域作为该时间段的热点位置. 热点位置与时间点结合就能形成对象一天的运动轨迹,根据对象运动轨迹中所对应的时间和位置就能进行对象位置预测.

收稿日期:2016 - 10 - 27;**修订日期:**2017 - 01 - 10

基金项目:国家高技术研究发展计划(863 计划)资助项目(2013AA014702)

作者简介:胡燕(1982—),女,湖北武汉人,北京邮电大学工程师,主要研究方向为信息处理、大数据.

1 K-Means 算法及相关定义

1.1 K-Means 算法

K-Means 算法是用于解决聚类问题的经典算法,是在给定样本点集合和 k 值的前提下进行聚类的简单算法. 假设 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$ 是要进行聚类的样本点,$\boldsymbol{x}_i$ 是其中任意一个样本. K-Means 算法将样本点划分为 k 个聚类,定义为 $\boldsymbol{C}=(\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_k)$,其主要思路是为每个聚类定义 c_i 个质心点. 首先,在样本点中随机选取 k 个质心点,然后计算样本点到该质心点的距离. 质心点到样本点的距离可以定义为:$\| \boldsymbol{x}_i-\boldsymbol{c}_j \|^2 (i\in n, j\in k)$. 计算每个样本点到质心点的距离,将该样本点归到距离最近的质心点类里,重复直到所有样本点都被归类.

然后根据上一步聚类结果重新计算 k 个质心点,重复之前的步骤将所有样本点重新归类,反复循环直到质心点没有变化. 以上过程可以表述为 $J=\sum_{j=1}^{k}\sum_{i=1}^{n}\| \boldsymbol{x}_i^{(j)}-\boldsymbol{c}_j \|^2$,式中:$\| \boldsymbol{x}_i^{(j)}-\boldsymbol{c}_j \|^2$ 是样本带点 $\boldsymbol{x}_i^{(j)}$ 到质心点 $\boldsymbol{c}_j$ 的距离,也就是到他们各自的聚类中心的距离.

1.2 相关定义

基于 K-Means 和时间匹配的位置预测模型与对象的历史位置信息、时间段、对象位置轨迹、预测准确率等概念相关,因此对以上概念做如下定义.

定义 1 位置点. 假设 $\boldsymbol{L}=(\boldsymbol{L}_1,\boldsymbol{L}_2,\cdots,\boldsymbol{L}_n)$ 是所有对象的历史位置点集合,对于任意的 L_i,其经纬度的位置信息可以记录为 $\boldsymbol{L}_i=(x_i,y_i)$.

定义 2 对象活动区域. 聚类的结果记为 $Cluster=(\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_k)$,其中 k 是给定的输入值. 聚类中每个位置点的最小和最大经度纬度形成运动区域,记录为:

$$O=(X_{\min},Y_{\min},X_{\max},Y_{\max}).$$

定义 3 时间段. 将一天 24 小时划分为多个不同的时间段,描述为:$T=([t_i,t_j],[t_j,t_k],\cdots,[t_n,t_i])(0\leqslant t_i<t_j<t_k<t_n\leqslant 24)$.

定义 4 热点位置. 对象出现在某个历史位点的概率可以描述为:

$$p_i=\frac{\boldsymbol{L}_i}{\sum_{i=j}^{n}\boldsymbol{L}_j}(1\leqslant i\leqslant n).$$

在 $T_{ij}=[t_i,t_j]$ 时间段内,对象出现概率最大的位置点为热点位置,记为 $M\mathrm{p}$,

$$M\mathrm{p}=\arg\max\frac{\boldsymbol{L}_i}{\sum_{i=j}^{n}\boldsymbol{L}_j}(1\leqslant i\leqslant n).$$

定义 5 对象运动轨迹. 对象在一天的不同时间段内对应的位置序列称为对象位置轨迹,其中某天的运动轨迹为:

$$S=([T_1,\boldsymbol{L}_1],[T_2,\boldsymbol{L}_2],\cdots,[T_n,\boldsymbol{L}_n]).$$

定义 6 位置预测准确率. 以天为单位对对象位置进行预测,对象在每天的每个时间段,热点位置实际发生的位置与运动轨迹中预测的运动区域相同则认为预测准确. 根据定义 5,将一天时间划分为 n 个时间段,其中有 m 个时间段的热点位置预测准确,位置预测准确率记为:

$$p_{\mathrm{a}}=\frac{O_m}{\sum_{i=1}^{n}O_n}(1\leqslant i\leqslant n, m\leqslant n).$$

2 基于 K-Means 和时间匹配的位置预测模型

笔者提出一种基于 K-Means 和时间匹配的位置预测模型,预测过程主要分两个阶段. 第一阶段为轨迹建模阶段,使用 K-Means 算法对历史访问位置点按照不同时间段进行聚类形式轨迹簇,计算出每个时间段出现最频繁的位置点作为该时间段的热点位置,形成对象每天的运动轨迹. 第一阶段中用 K-Means 对历史位置点进行聚类,形成对象运动区域,目的是提高预测的准确性,同时对运动区域的预测不影响移动服务的推送和推荐. 第二阶段根据时间匹配原则对位置进行预测.

2.1 轨迹建模

轨迹建模阶段是对历史位置点进行聚类,形成对象运动区域. 使用任意两个历史位置点之间的距离用 K-Means 算法进行聚类,其距离表示为 $D_{ij}=\| x_i-y_i \|^2$,聚类过程如下:

输入:所有历史位置点 $\boldsymbol{L}_i=(x_i,y_i)$ 和 k 值.

输出:聚类结果.

①在历史位置点中随机选择 k 个样本点为初始质心.

②计算某个历史位置点到每个质心点的距离,将该历史位置点归入距离最近的那个质心点. 历史位置点到所有质心点的最小距离可以表示为:

$$D_{ik}=\arg\min\sum_{i=1}^{k}\sum_{x_i\in L_i}\| \boldsymbol{x}_i-\boldsymbol{m}_k \|^2.$$

③当所有的历史位置点被分配以后,重新计算质心的位置.

④重复②和③直到质心点没有变化,按照计算

最短距离的方式将所有的历史位置点进行聚类.

聚类的结果记为 $Cluster = (\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k)$,其中 k 是给定的输入值. 聚类完成后计算对象活动区域,任意一个 c_i 的对象位置区域为:

$$O = (X_{i\min}, Y_{i\min}, X_{i\max}, Y_{i\max}).$$

由于每个对象出现的历史位置点都有对应的时间,因此对象出现时间与对象位置预测是有紧密联系的. 假设 $O_i = (O_a, O_b, \cdots, O_k)$ 为对象在时间段 $T_i = [T_{ij}, T_{jk}, \cdots, T_{ni}]$ 的热点位置,按照时间序列可以形成对象一天的历史轨迹,即

$$S = \{\langle T_{ij}, O_a\rangle, \langle T_{jk}, O_b\rangle, \cdots, \langle T_{ni}, O_k\rangle\},$$
$$(0 \leqslant t_i < t_j \cdots, t_n \leqslant 24, 0 < a, b, \cdots, < k).$$

2.2　轨迹预测

通过轨迹建模阶段对对象位置数据进行处理,使用轨迹建模算法得到的对象运动轨迹,该模型由不同时间序列组成,每个时间段的热点位置用概率估算的方法对对象未来的运动轨迹进行预测.

$$S' = \{\langle T_{ij}, O_a\rangle, \langle T_{jk}, O_b\rangle, \cdots, \langle T_{ni}, O_k\rangle\},$$
$$(0 \leqslant t_i < t_j, \cdots, t_n \leqslant 24, 0 < a, b, \cdots < k).$$

3　实验和结果

实验使用加利佛尼亚大学发布的真实数据,这些数据是从校园 AP 获取的对象位置信息. 数据包含 AP 信息和对象访问信息. AP 信息包含 AP 的经纬度、高度等信息. 对象访问信息是通过对象的终设备软件采集,当对象打开该软件时软件将会记录访问 AP 的信息,包括 AP 编号、访问时间、信号强度等信息. 笔者从发布的数据中选取 50 个对象两个星期的访问信息作为历史位置信息,其中前一个星期数据用于轨迹建模,后一个星期的访问信息用于轨迹更新.

首先将所有历史记录中出现的 AP 使用 K-Means 聚类,假设 $k = 8$,所有的 AP 按照位置聚类为 8 个类,图 1 为聚类结果.

根据作息规律将一天 24 小时划分为 6 个时间段,分别是 t1 = [0am,6am]、t2 = [6am,9am]、t3 = [9am, 12am]、t4 = [12am, 14pm]、t5 = [14pm, 18pm]、t6 = [18pm,24pm]. 然后计算每个时间段访问次数最多的位置作为该时间段的代表位置. 这样就能按照时间段形成对象的位置轨迹,并利用该对象一天的运动轨迹对对象位置进行预测.

实验选取 50 个采集信息较多的对象位置进行实验,图 2 为不同时间段预测准确率的对比结果. 在 t1 时间段,两种方式的预测准确率相同,因为该时间段([0am, 6am])是对象睡觉时间,因此聚类前后准确率无变化. t2([6am, 9am])时间段是早晨活动时间,该时间对象的活动范围不太固定,聚类后的预测准确率低于聚类前. t3 ~ t6 时间为学习和晚间活动时间,该时间段对象活动范围相对固定,因此聚类后预测准确率高于聚类前. 实验结果证明校园对象在一定时间段内在固定区域内活动较多,笔者提出的基于 K-Means 和时间匹配的位置预测模型适合针对校园对象进行位置预测. 使用模型前按照时间段对位置点预测,不针对运动区域进行预测,图 3 显示使用预测模型前的预测准确率为 39.7%,使用预测模型后的预测准确率为 60.3%.

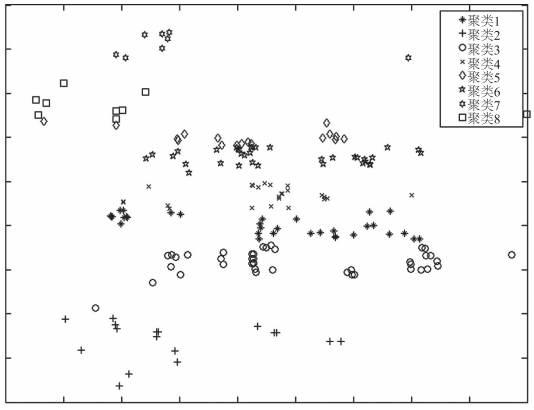

图 1　K-Means 算法聚类结果

Fig. 1　Cluster result with K-Means algorithm

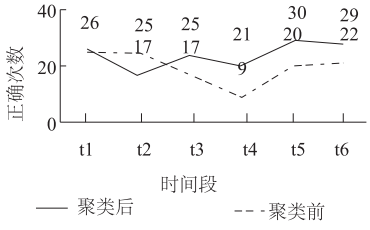

图 2　不同时间段预测准确率对比结果

Fig. 2　Results of comparison in different time segments

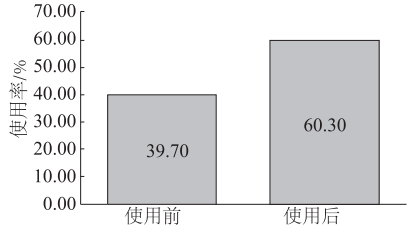

图 3　使用预测模型前和使用预测模型后预测准确率对比结果

Fig. 3　Prediction ratio before and after using model

4　结论

提出一种基于 K-Means 和时间匹配的位置预

测模型,该模型使用 K-Means 算法对历史位置点信息聚类,形成对象运动区域.在此基础上笔者又提出轨迹建模算法用于确定对象历史运动轨迹,并根据时间匹配原则按天对对象轨迹进行预测.为了实现该模型,使用加利佛利亚大学的真实数据进行试验.试验证明,该模型提高了位置预测的准确性.因此,基于 K-Means 和时间匹配的位置预测模型是较为准确、高效的.

参考文献:

[1] 李雯,夏士雄,刘峰,等. 基于运动趋势的移动对象位置预测[J]. 通信学报, 2014, 35(2): 46-53.

[2] 于瑞云,夏兴有,李婕,等. 参与式感知系统中基于社会关系的移动对象位置预测算法[J]. 计算机学报, 2015, 38(2): 374-385.

[3] ROBARDS W M, SUNEHAG P. Semi-markov K-Means clustering and activity recognition from body-worn sensors[C]. 2009 Ninth IEEE International Conference on Data Mining. 2009.

[4] YANG Y, WANG Z L, ZHANG Q, et al. A time based markov model for automatic position-dependent services in smart Home[C]. 2010 Chinese Control and Decision Conference. Beijing, 2010.

[5] 赵杨,田国会,尹建芹,等. 家庭智能空间下基于HMM的人轨迹分析方法[J]. 模式识别与人工智能,2015,28(6): 542-549.

[6] 彭曲,丁治明,郭黎敏. 基于马尔可夫链的轨迹预测[J]. 计算机科学,2010,37(8):189-192.

[7] 李婕,夏兴有,王兴伟,等. 机会认知网络中基于社会关系的节点位置预测算法[J]. 东北大学学报(自然科学版),2014, 35(12):1701-1705.

[8] LI W, XIA S X, LIU F, et al. Markov location prediction algorithm based on dynamic social ties[C]. IEICE TRANS. INF. & SYST,2015.

[9] 赵雪涵. 基于密度聚类的用户轨迹预测算法研究[D]. 西安:西安理工大学计算机等院. 2014.

[10] FULOP P, SZABO S, SZALKA T. Accuracy of random walk and markovian mobility models in location prediction methods[C]. 15th International Conference on Software, Telecommunications and Computer Networks. 2007.

[11] RACHURI K, MURTHY R C. Level biased random walk for information discovery in wireless sensor networks[C]. IEEE International Conference on Communications. 2009.

[12] XU J H, LIU H. Web user clustering analysis based on K-Means algorithm[C]. International Conference on Information Networking and Automation (ICINA). 2010.

Location Prediction Model Based on K-Means Algorithm and Time Matching

HU Yan, ZHU Xiaoying, MA Gang

(Network and Information Center, Beijing University of Posts and Communications, Beijing 100786, China)

Abstract: Location prediction was critical to mobile service because various kinds of applications were tightly combined with user's location. However, location prediction was a challenging work because location capturing was always not continuous and user's behavior were uncertain and irregular. To improve the location prediction accuracy rate, this paper proposed a location prediction model based on K-Means algorithm and time matching. For the mobile service always region oriented, we first clusted history location using K-Means algorithm to define several regions. Then we divided every day time into several segments and calculated the maximum probability location in every time segment. A trajectory of a user in one day was formed with trajectory model and trajectory updating model which proposed in this paper. We could predict user' location with time matching method. At last, we did experiments with real location data in campus which captured by APs. The prediction out come with K-Means was compared to the outcome without model based on K-Means algorithm. The experiment result shows that accuracy rate of our model was higher than the prediction without new model. So, more location services could be provided to users with this new model.

Key words: location prediction; K-Means algorithm; time matching; cluster