基于改进LCSS的移动用户轨迹相似性查询算法研究

陈少权

（广州杰赛科技股份有限公司，广东 广州 510310）

【摘　要】　为了解决由于移动用户轨迹数据具有随机性和繁杂性导致算法效率和精度低的问题，首先抽取用户轨迹时间位置序列，然后基于用户的逗留时长采用加权FP树挖掘移动用户的常驻区域以解决用户轨迹的随机性，最后提出结合用户出行的时间和地理因素的LCSS算法衡量用户轨迹相似性。实验证明，该算法具有一定的有效性和扩展性。

【关键词】　轨迹相似性　FP树　最长公共子序列　时间相似性系数

doi:10.3969/j.issn.1006-1010.2017.06.014　中图分类号：TP391.4　文献标志码：A　文章编号：1006-1010(2017)06-0077-06

引用格式：陈少权. 基于改进LCSS的移动用户轨迹相似性查询算法研究[J]. 移动通信, 2017,41(6): 77-82.

Research on Query Algorithm of Mobile User's Trajectory Similarity Based on Improved LCSS

CHEN Shao-quan

(GCI Science & Technology Co., Ltd., Guangzhou 510310, China)

[Abstract]　In order to deal with the algorithm efficiency and precision due to the randomness and complexity of mobile user's trajectory data, the time location sequence of user trajectory was extracted. Then, the weighted FP tree was used to mine mobile user's resident region based on user's residence time to cope with the randomness of user's trajectory. Finally, LCSS algorithm combined user's traveling time with geographical factor was proposed to evaluate the similarity of user's trajectory. Experiments show that the proposed algorithm has satisfactory effectiveness and expansion.

[Key words]　trajectory similarity　FP tree　longest common subsequence　time similarity coefficient

1　引言

随着移动通信和移动应用的快速发展，用户对手机的使用率及依赖性不断提高，移动运营商积累了大量移动用户实时记录的定位数据。分析移动用户位置的相似性，提取移动用户的相似路径在出行路径预测、兴趣区域发现、轨迹聚类、个性化路径推荐等领域具有广泛的应用。如何利用丰富的移动用户定位数据找到合适轨迹的表示方法，如何高效计算移动用户轨迹间的相似性已经成为工业界和学术界的研究热点[1]。贾振美针对移动对象的稀疏数据处理的问题，提出了稀疏数据处理的具体方法，并结合时间因素挖掘用户的频繁轨迹模式，利用相似用户聚类方法实现同类用户位置的预测[2]。张用川通过对手机用户移动轨迹数据的语义环境信息进行研究，创新性提出地理位置语义化的方法，挖掘用户出行模[3]。罗家顺通过预先对用户位置进行定义建立用户－位置信息模型，然后结合时间效应利用协同过滤的算法找到区域性相似的用户，实现实时的多维动态

收稿日期：2016-11-28

责任编辑：刘妙　liumiao@mbcom.cn

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

用户兴趣推荐[4]。肖啸骐针对移动对象轨迹在时间维度分布不均匀的特点，提出了一种基于关键点和时间分段的稀疏轨迹相似性的方法，提升轨迹相似性度量的运行效率和准确度[5]。吕瑞鹏提出一种基于移动概括的相似度方法来衡量用户移动轨迹的相似度[6]。本文在现有研究者的研究成果基础上，从移动用户的原始轨迹数据抽取位置序列，同时将位置序列映射为具有时间和地理位置信息的序列，解决由于移动用户轨迹数据的稀疏性导致相似度算法效率低下的问题。再结合用户逗留的时长，通过FP-tree（Frequent Pattern-tree，频繁模式树）的加权频繁模式挖掘移动用户轨迹的频繁序列，解决由于用户轨迹随机性和繁杂性而导致算法效率低下的问题。最后通过改进LCSS（Longest Common Subsequence，最长公共子序列）的方法，结合时间和地理因素衡量用户轨迹的相似性。

2 移动用户轨迹表示及相似度的研究

2.1 移动用户轨迹表示

原始的移动用户轨迹数据一般由离散的数据组成，由于信号或者用户发生业务的原因，这些用户的轨迹数据会出现分布不均匀、不连续的特点。为了有效挖掘移动用户的行为，不少学者通过对原始的移动用户数据做了相关的处理，得到有效的用户轨迹数据。常见的移动用户轨迹表现方法有：基于FP-Growth算法对用户的常驻地模式进行挖掘，得到若干个有效的用户轨迹数据；基于位置序列的抽取，将不同用户的位置映射为不同的字符串；利用最小包围盒技术描述移动用户的停留区域的基于停留区轨迹表示方法；利用手机轨迹数据的出行目的为目标，采用语义技术来提取用户出行的地理位置。

2.2 相似度衡量的方法

衡量相似度的方法有很多，常见的有欧式距离、动态时间规整（Dynamic Time Warping，DTW）、编辑距离（Edit Distance on Real Sequence，EDR）、最长公共子序列（Longest Common Subsequence，LCSS）、最大时间出现法（Maximum Co-occurrence Time，MCT）、余弦相似性（Cosine Similarity）、Hausdorff距离等，这些方法的适用数据类型和应用场景不尽相同，因此在选择时需要充分考虑应用情况[7-9]。基于本文轨迹数据的特点，以下重点介绍欧式距离、DTW、LCSS三种衡量相似度的算法。

（1）欧式距离

欧式距离是通过计算每个时间点上轨迹所对应的两个点的欧式距离，然后再对所有点的欧式距离进行综合处理，总和处理包括平均值、求和、取中值等方式。

$$EU=\sum_{k=1}^{n}\mathrm{dist}(p_k^A,p_k^B) \quad (1)$$

$$\mathrm{dist}(p_k^A,p_k^B)=\sqrt{(p_{k,x}^A-p_{k,x}^B)^2+(p_{k,y}^A-p_{k,y}^B)^2} \quad (2)$$

其中，EU表示欧式距离，$\mathrm{dist}(p_k^A,p_k^B)$表示用户$A$和$B$在某段时间段内的距离；$p_k^A$、$p_k^B$表示$A$和$B$在$k$时刻的位置；$p_{k,x}^A$、$p_{k,x}^B$表示用户$A$和用户$B$在$x$维度的位置，同理，$p_{k,y}^A$、$p_{k,y}^B$表示用户$A$和用户$B$在$y$维度的位置。

欧式距离的缺点是容易受到噪音的影响，特别是现实中两个移动用户的轨迹在时间和个数上都存在很大差异，因此采用该方法的时候必须对移动用户的轨迹数据进行预处理。

（2）动态时间规划

动态规划方法解决了欧式距离对采样过于苛刻的要求，采用重复点之前的记录点填补对应空缺的方式，以求出的最小距离作为轨迹的相似性度量[11]。

假设有两个用户轨迹空间域的离散采样$P=<p_1, p_2, \cdots, p_m>$和$Q=<q_1, q_2, \cdots, q_m>$，基于DTW对两条轨迹的采样点数量没有任何的要求，那么两条轨迹之间的相似度公式[12]为：

$$\mathrm{DTW}(P,Q)=f(m,n) \quad (3)$$

$$f(i,j)=\|p_i-q_j\|+\min\begin{cases}f(i,j-1)\\f(i-1,j)\\f(i-1,j-1)\end{cases} \quad (4)$$

式中，||.||为两点坐标的二范数，也就是两点之间的欧式距离。

（3）最长公共子序列（LCSS）

DTW和欧式距离对轨迹的个别点差异性非常敏感，如果两个时间序列在大多数时间段具有相似的形态，仅仅在很短的时间具有一定的差异，那么欧式距离和DTW无法准确衡量这两个时间序列的相似度。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

LCSS很好地解决了这些问题。

假设有两条长度分别为n和m的时间序列数据A和B，那么最长公共子序列的长度为[13]：

$$\mathrm{LCSS}(A,B)=\begin{cases}0, if\ \ A=\varnothing or B=\varnothing;\\ 1+\mathrm{LCSS}(a_{t-1},b_{t-1}),\ if(\mathrm{dist}(a_t,b_i)<\gamma);\\ \max(\mathrm{LCSS}(a_{t-1},b_i),\ \mathrm{LCSS}(a_t,b_{t-1})),\ otherwise\end{cases} \quad (5)$$

其中，γ为一个成员相似阈值，t=1, 2, …, n；i=1, 2, …, m。基于上述公式，基于公共子序列的相似度公式为[14]：

$$D_{\mathrm{LCSS}}=1-(\mathrm{LCSS}(A,B))/\min(len_A,len_B) \quad (6)$$

3 基于改进LCSS的移动用户轨迹相似性思路分析

基于改进LCSS的移动用户轨迹相似性查询算法的研究包括几个重要的步骤：

（1）抽取位置序列，将位置序列映射为具有时间和地理位置信息的序列，以发生时间的序列表示移动用户的轨迹。

（2）采用FP-Growth算法挖掘移动用户轨迹的频繁序列。

（3）结合时间和地理因素，采用改进LCSS的方法衡量用户轨迹的相似性。

3.1 移动用户轨迹的表示

移动用户的轨迹一般由一系列按照时间依次排序的位置组成[10]，$Tr_i=\{(L_1, t_1), (L_2, t_2), \cdots, (L_i, t_i), \cdots, (L_n, t_n)\}$。$(L_i, t_i)$表示用户出现在某个基站的位置$L_i$对应的时间$t_i$。

移动用户轨迹是按照时间序列形成有序的集合，因此，在考虑时间因素的情况下，可通过移动用户的轨迹抽取移动用户的时间位置序列。上述的移动用户轨迹可表示为$Tr_i=\{(L_1, L_2, t_1, t_2), (L_2, L_3, t_2, t_3), \cdots, (L_i, t_i, L_{i+1}, t_{i+1}), \cdots, (L_{n-1}, t_{n-1}, L_n, t_n)\}$。序列中$(L_1, L_2, t_1, t_2)$表示移动用户在时刻$t_1$出现在基站$L_1$，然后在时刻$t_2$离开基站$L_1$前往基站$L_2$。

3.2 移动用户轨迹频繁序列的挖掘

对于移动用户轨迹数据的频繁模式定义为如下形式：

$$L_i \rightarrow L_j \quad (7)$$

公式（7）的定义是一个移动用户从位置L_i向位置L_j移动的规律。移动用户频繁轨迹提取是从移动用户移动轨迹数据集中提取支持度大于最小支持度阈值的集合。因此，移动用户频繁模式反映了移动用户群体在移动行为上具有相同特征或是相同规律[15]。但由于频繁项集在运算过程中需要付出更大的代价，因此本文引入闭合频繁项集来保证挖掘得到的移动用户行为信息量最全面且数据规模最小。Pasquier于1999年提出频繁闭合项集的概念，定义了频繁闭合移动模式。假设频繁移动模式Tp_i属于频繁闭合移动模式，其必须满足：在频繁模式集中不存在任一个模式Tp_j，满足$Tp_j \supseteq Tp_i$，且$\mathrm{support}(Tp_j) \geqslant Tp_i$。由于考虑到移动用户在基站的逗留时间，本文以频繁闭合序列模式挖掘经典算法，以基站平均逗留时间作为项目权重，以各项目count值降序依次为头节点和其他节点，生成条件模式基，然后采用条件模式基构造对应的加权条件FP树，最后并按照设定加权支持度的阈值判断相应的频繁模式。

3.3 基于改进LCSS的移动用户轨迹相似性查询算法

为了提高移动用户轨迹识别的准确性，在通过TP树获得用户常驻区域模式的基础上，结合时间因素，以时间系数反映所有用户在邻近时间在相同的地理位置的比例。时间相似性系数的公式为：

$$COL=\frac{\sum_{i=1}^{n(u)}\sum_{j=1}^{n(v)}(\Delta T-|T_i(u)-T_j(v)|)\delta(L_i(u),L_j(v))}{\sum_{i=1}^{n(u)}\sum_{j=1}^{n(v)}(\Delta T-|T_i(u)-T_j(v)|)} \quad (8)$$

其中，ΔT为精度（一般设为1个小时），$T_i(u)$表示移动用户u在某一个时间精度内达到某一个基站$L_i(u)$的时刻，$T_j(v)$表示移动用户v在某一个时间精度内达到某一个基站$L_j(v)$的时刻，$\delta(L_i(u), L_j(v))$是一个重合性公式，当两个用户的基站重合时，值为1，否则值为0。

结合时间因素，改进的LCSS的相似度算法为：

$$D_{\mathrm{LCSS}}=\frac{1-(\mathrm{LCSS}(u,v))}{\min(len_u,len_v)}\times\frac{\sum_{i=1}^{n(u)}\sum_{j=1}^{n(v)}(\Delta T-|T_i(u)-T_j(v)|)\delta(L_i(u),L_j(v))}{\sum_{i=1}^{n(u)}\sum_{j=1}^{n(v)}(\Delta T-|T_i(u)-T_j(v)|)} \quad (9)$$

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

公式的第一部分表示用户u和用户v一天的最长公共子序列，第二部分表示在每一个时间精度下，两位用户在邻近时间在相同的地理位置的比例。

4 实验分析

4.1 用户移动轨迹数据的提取和预处理

本次实验随机抽取某运营商的10 000名移动用户两周的轨迹数据，包括用户的发生业务的起始时间、起始基站名称、切换基站时间、切换基站名称、在每一个基站的逗留时长、主叫号码、被叫号码、用户发生的业务类型等。在对数据进行挖掘之前，先对数据进行预处理，剔除与求解轨迹相似度无关的字段，然后抽取用户的时间位置序列，最后按照发生业务的起始时间的顺序排列每一个用户的时间位置数据。具体如表1所示。

表1　用户轨迹预处理结果

移动用户号码	起始时间	结束时间	起始基站CI	结束基站CI
18676445***	20140601001905	20140601001918	2353	672
18676445***	20140601001918	20140601001932	672	6582
18676445***	20140601001932	20140601001942	6582	31058
18676445***	20140601001942	20140601001745	42487	31271
18676445***	20140601001948	20140601002008	31271	57522
18676445***	20140602001017	20140602001140	57522	57523
18676445***	20140602001140	20140602011351	57523	57522
18676445***	20140602001351	20140602031846	57522	57522
18676445***	20140602001446	20140602001846	57522	57522

从表1可以看出，经过对移动用户原始的轨迹进行预处理之后，得到每一个移动用户的时间位置信息，为下一步数据挖掘做准备。

4.2 采用FP树挖掘移动用户轨迹频繁序列

（1）对用户移动轨迹的项目以及项集的数据处理

在获取用户时间位置信息的基础上，计算移动用户在每一个基站的平均逗留时间，以此作为项目权重。项目名称及权重如表2所示：

表2　项目名称及权重

项目名称 —— 基站ID	权重 —— 平均逗留时间/s
2353	286
672	30
6582	45
42487	67
31271	15
57522	266

从用户移动轨迹处理结果提取用户的项集X={2353-672-6582-42487-31271-57522}。根据用户在每一个基站的逗留时间设置每一个项目（基站）的权重，当项目（基站）具有一个权重后，用户发生轨迹的项目项集（基站组合）的权重定义为各项目权重的平均值[17]。例如表3中，X={2353-42487-672-6582}，用户的移动轨迹项集权重为WT(t)=（286+67+30+45）/4=107，经过归一化操作之后，该项集的归一化权重为0.1488。

表3　用户轨迹项集及权重

项集（用户经历的基站组合）	用户移动轨迹项集权重（项集平均逗留时间）	均一化的用户移动权重
2353-42487-672-6582	107	0.1488
2353-42487	176.5	0.2454
42487-672	48.5	0.0674
2353-42487-672-31271	99.5	0.1384
2353-42487-672-6582-31271	88.6	0.1232
57522-6582	199	0.2767

（2）建立加权FP树

由表3可得到各项目{2353，672，6582，42487，31271，57522}的count为{5，4，3，4，2，1}。结合用户在每一个基站的逗留时长，根据FP树构造的思想，得到某用户移动轨迹的加权FP树。基于用户在基站逗留时长的加权FP树如图1所示。

根据加权FP树导出用户逗留基站的权重分别是：2353:0.6558；42487:0.6558；6582:0.5487；672:0.3394；31271:0.2616；57522:0.2767。

设最小支持度Wminsup=0.45，那么根据上述的加权条件树得出的频繁模式如表4所示。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

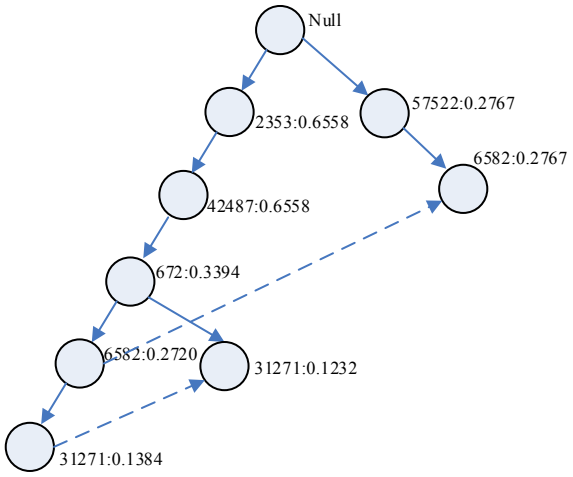

图1　基于用户在基站逗留时长的加权FP树

表4　加权条件FP树和频繁模式

基站 ID	加权条件 FP 树	频繁模式
42487	<2553:0.6558>	2553-42486:0.6558
672	<2353:0.4014, 42487:0.4778>	2353-672:0.4014, 42487-672:0.4778, 2353-42487-672:0.4778
6582	<2353:0.2720, 42487:0.2720, 6727:0.2720> <57522:0.2767>	-
31271	<2353:0.2612, 42487:0.2612, 6727:0.2612, 6582:0.2612>	-

4.3 基于LCSS算法评价移动用户轨迹相似性的结果

基于加权FP树提取移动用户的常驻地点，再结合移动用户在常驻地点的时间因素，采用公式（9）计算的10 000名移动用户工作日的轨迹相似度的结果如表5所示：

表5　基于LCSS算法评价移动用户轨迹相似性的准确性

LCSS区间	用户数	准确率/%
≥0.7	10 000	65.17
≥0.6	10 000	79.73
≥0.5	10 000	88.56
≥0.4	10 000	91.17

注：在计算轨迹相似性时会剔除电话号码的字段进行相似性计算，然后根据相似性的结果再关联电话号码进行正确率的验证。

从实验可知，LCSS区间的合理范围在（0.4，0.5），通信运营商或者移动运营商可根据不同的业务需求挖掘不同用户之间轨迹的相似性，为营销工作提供数据支撑。

5　结束语

随着移动互联网的快速发展，大数据洪流已经全方位深入到移动用户的生活中。如何有效利用移动用户的海量数据为电信运营商、移动运营商或者其他商家提供有效的营销数据支撑已经成为研究的热点。本文首先按照一定的规则对移动用户轨迹数据进行时间和位置序列预处理，然后采用FP树挖掘用户的常驻地点，最后通过改进的LCSS算法来判断移动用户轨迹的相似性。实验证明，该算法具有较高的准确率和扩展性。下一步的工作是设计分布式算法，以支持大规模的移动用户轨迹相似性的计算，提升计算的速度。

参考文献：

[1] 裴剑,彭敦陆. 一种基于LCSS的相似车辆轨迹查找方法[J]. 小型微型计算机系统, 2016(6): 1197-1202.

[2] 贾振美. 面向稀疏轨迹数据的位置预测方法研究[D]. 沈阳: 东北大学, 2014.

[3] 张用川. 基于手机定位数据的用户出行规律分析[D]. 昆明: 昆明理工大学, 2013.

[4] 罗家顺. 面向移动用户的协同过滤推荐算法研究[D]. 长春: 吉林大学, 2016.

[5] 肖啸骐. 一个有效的稀疏轨迹数据相似性度量[J]. 微型电脑应用, 2014(4): 25-30.

[6] 吕瑞鹏. 基于移动概括的新用户相似度衡量方法[D]. 济南: 山东大学, 2014.

[7] Chen L, Ozsu M T, Ora V. Robust and fast similarity search for moving object trajectories[A]. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data[C]. ACM, 2005: 491-501.

[8] Lee J G, Han J, Whang K Y. Trajectory clustering: a partition and group framework[A]. Proceeding of 2007 ACM SIGMOD International Conference on Management of Data[C]. ACM, 2007: 593-604.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[9] Lee S L, Chun S J, Kim D H, et al. Similarity search for multidimensional data sequences[A]. Data Engineering, Proceedings 16th International Conference on IEEE[C]. 2000: 599-608.
[10] 肖艳丽. 基于位置序列的广义后缀树用户相似性计算方法[J]. 计算机应用, 2015,35(6): 1654-1658.
[11] 郭岩,罗珞珈,汪洋,等. 一种基于DTW改进的轨迹相似度算法[J]. 研究与开发, 2016,35(9): 66-71.
[12] SAKURAI Y，YOSHIKAWA M，FALOUTSOS C. FTW: fast similarity search under the time warping distance[A]. Symposium on Principles of Database System[C]. 2005: 491-502.
[13] Marascu A, Khan S A, Palpanas T. Scalable similarity matching in streaming time series[A]. PAKDD´12 Proceedings of the 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining[C]. 2012: 218-230.
[14] Vlachos M, Kollios G, Gunopulos D. Discovering similar multidimensional trajectories[A]. International Conference on Data Engineering[C]. 2002: 673-684.
[15] 王亮,汪梅,郭鑫颖,等. 面向移动时空轨迹数据的频繁闭合模式挖掘[J]. 西安科技大学学报, 2016,36(4): 573-576.
[16] Pasquier N, Bastide Y, Taouil R, et al. Discovering frequent closed itemsets for association rules[A]. ICDT´99 Proceedings of the 7th International Conference on Database Theory[C]. 1999: 398-416.
[17] 陈文. 基于FP树的加权频繁模式挖掘算法[J]. 计算机工程, 2012,38(6): 63-65.★

作者简介

陈少权：中级工程师，学士毕业于南昌大学电子与信息系统专业，现任职于广州杰赛科技股份有限公司通信规划设计院副总经理，目前主要从事公司运营管理研究工作。

（上接第 76 页）

问题查找算法[J]. 移动通信, 2011,35(8): 18-22.
[5] 牟大维,甘小莺,徐友云,等. 基于扇区化的OFDM小区选择算法[J]. 电讯技术, 2006,46(5): 79-83.
[6] 王国民,雷萍,邱恺. 地域通信网干线节点抗干扰等效推算探析[J]. 电子信息对抗技术, 2015,30(5): 63-66.
[7] 李丽,宋燕辉,朱江军. 基于功率控制的LTE系统下行ICIC算法研究[J]. 电视技术, 2013,37(23): 167-170.
[8] 李斌,朱宇霞. LTE系统中切换优化算法的研究[J]. 电视技术, 2013,37(3): 109-112.★

作者简介

叶蔼笙：硕士毕业于华南理工大学，现任中国电信股份有限公司广州分公司无线优化中心网络优化室副经理，具有七年无线网络优化经验，主要从事LTE性能分析、系统参数优化等工作。

史俊辉：学士毕业于南京邮电学院，现任中国电信股份有限公司广州分公司无线优化中心网络优化室经理，具有十六年无线网络优化经验，主要从事LTE系统参数优化、项目管理等工作。

曹家燕：硕士毕业于中国科学技术大学，现任中国电信股份有限公司广州分公司无线优化中心常规分析工程师，具有两年无线网络优化经验，主要从事LTE性能分析、数据挖掘等工作。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net