基于轨迹数据和并行处理技术的兴趣区域推荐

姓名	学号	初步分工	联系信息
吴军	MF1933101	程序实现	18256507929
吴华清	MF1933100	选题+程序实现	18094241231
甘世维	MG1933015	程序实现	15675102866
王宁	DZ1933027	程序实现	18013100213

1. 问题背景

随着各种定位技术（如全球定位系统 GPS 和无线蜂窝网）的发展和普及，用户可以很方便的获取个人位置信息，使用各种基于位置的服务（LBS），并将自己的移动过程以轨迹的形式记录下来。其实，轨迹记录了用户在真实世界的活动，而这些活动将在一定程度上体现了个人的意图、喜好和行为模式。如某个用户的轨迹经常出现在运动场馆，表明该用户可能会喜欢体育活动；而经常穿越湖光山色的路线也表征用户对户外活动的喜好。更细粒度的分析甚至可以根据用户经常光顾的餐馆类别（如川菜馆、湘菜馆）来判别出用户的口味。 因此，如何挖掘轨迹中蕴含的知识就变得尤其重要[1][1][1]。

2. 技术难点

随着 GPS 技术以及各种移动终端的普及，将产生大量 GB,甚至 TB 级别的 GPS 轨迹数据，比如，北京拥有约 7 万辆出租车，为了管理和调度这些车辆，每辆车都配备了 GPS 装置。如果每辆车每秒钟都向交通管理部门发送自己的坐标,一天的产生的数据量就将达到 1 个 TB，如果将发送频率减小到分钟级别，那么一天的总的数据量也会接近 GB 级别。因此，传统的集中式地处理数据的方式不能胜任日益增长的轨迹数据规模,需要使用大数据的并行处理技术来处理分析数据。

另外，由于用户的轨迹数据具有时间和空间的二维属性，这使得在使用一些传统的数据挖掘模型的时候，需要对轨迹数据进行一些相关的建模，从而适用于相关的算法模型。

3. 拟解决的问题

一条 GPS 轨迹通常由一系列带有时间戳的坐标点组成。每个坐标点包含了经度、纬度和海拔高度等基本信息。一个人在一段时间内的活动就可记录为这样一条连续的轨迹。在这条轨迹中，我们可以通过算法检测出一些用户停留过的地方。这个停留点并不是指速度为零的

点，而是由一组实际的 GPS 点构成，如图 1 中 p3, p4, p5 和 p6 构成了一个停留点 s。它表示用户在某个区域内滞留的时间超过了一定的时间范围。与其他 GPS 点相比，这些停留点含有更重要的语义信息，如用户去过的餐馆和电影院等。基于这些停留点，一个用户的历史轨迹就可以表达为一个停留点序列，如 s1->s2->…sn(中间间隔分别为△t1, △t2, △tn)。这个序列抓住了用户行为的重点，同时也大大减轻了数据处理量。

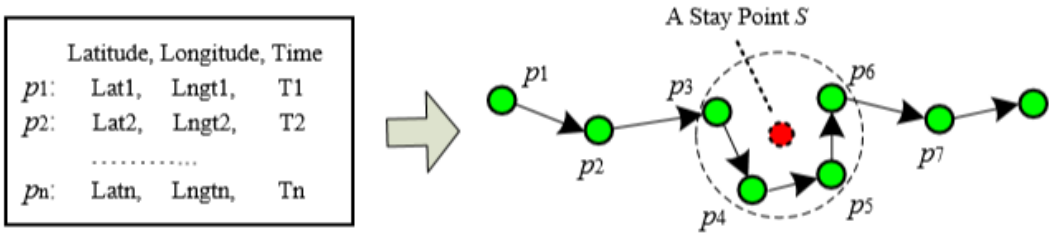

图 1. 一条 GPS 轨迹样例

由于用户多次访问同一地点所产生的停留点并不完全一致（坐标会有偏差），直接对停留点进行比较并不可行。因此，我们需要对从轨迹中提取出来的停留点进行聚类。这样相近的停留点就会被分配到同一个聚类中。此后，我们再用各个停留点所归属的聚类来替换这个停留点，将停留点序列进一步转化为聚类的序列，如图 2 所示。这样用户在不同时间段的历史轨迹就可比了。

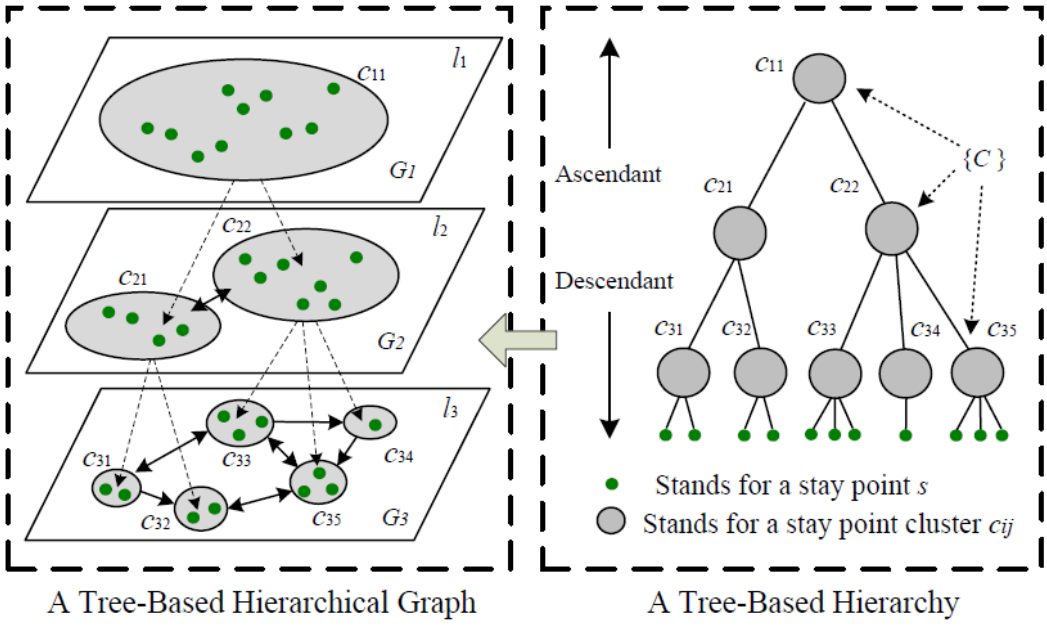

图 2. 基于层级图模型的多用户轨迹聚合

有了用户历史轨迹的模型，我们可以用多种算法（如 FP-growth、Closet+等）来挖掘

这个数据中的频繁项集。如用户 A 经常在周末早上去中关村、用户 A 经常在周五晚上去超市等。进一步，这些频繁模式，可以相互组合和连接，从而发现一些表征了用户生活、行为 规律的顺序模式（sequential pattern）。比如，通常用户 A 在周末早上会去中关村看电影，然后下午去西单买东西。

综上，我们要解决的问题就是，给定一个轨迹数据集，我们需要利用相关算法计算出每条轨迹的停留点，然后对于整个数据集上的所有的停留点进行层次聚类，从而利用聚类将轨迹数据转化为事务数据（一个聚类中的用户属于同一个事务的项集），在这个基础上，进行频繁项集等相关的数据挖掘，并且定义相关的相似用户，给相关的推荐算法给相似用户推荐兴趣区域[2]。

4. 基本解决方法和设计思路

由于每条轨迹数据的计算停留点的方式是独立的，所以可以利用 Hadoop 进行并行化的计算。之后进行层次聚类的时候，可以使用并行化的 k-Means 算法进行 MapReduce 的并行化计算；得到聚类之后，也就意味已经将轨迹数据转化为了事务数据，这样再利用相关的推荐算法进行给相似用户进行相关的推荐。

整个过程的步骤如下：

1. 给定一系列轨迹数据，使用 MapReduce 并行计算每一条轨迹的停留点，并且基于停留点，进一步计算一条轨迹的停留区域（这里可以设置成停留点的最小外接矩形或者最小外接圆）。
2. 对于每一条轨迹计算出的停留区域，通过建模转化为唯一标识的对象（例如可以使用停留区域最小外接圆的圆心和半径来唯一标识一个停留区域），进而对停留点进行层次聚类（例如使用 k-means 算法，可并行化实现）
3. 通过层次聚类，便将轨迹数据转化成了事务数据，同一个聚类的区域被划分在同一个项集中。我们通过最大频繁项集挖掘算法（可并行化实现）来定义相似用户，在同一个频繁项集中的用户被称为相似用户，其相似度为 1，否则为 0；
4. 有了相似用户的定义之后，我们使用协同过滤算法（可并行化实现）或者关联规则来推荐给用户 Top-n 个用户兴趣区域。

参考文献

[1] 郑宇, and 谢幸. ″基于用户轨迹挖掘的智能位置服务.″ 中国计算机学会通讯 6.6 (2010): 23-30.

[2] 龙玉绒,王丽珍,陈红梅.基于用户轨迹数据的用户兴趣区域推荐[J].软件工程,2019,22(11):8-14.

[3] Shad S A . 移动用户轨迹与行为模式挖掘方法研究[D]. 中国科学技术大学, 2013.

[4] 胡立, 陈健, 沈书毅, et al. 基于用户轨迹聚类分析的推荐算法研究[C]// 第 29 届中国数据库学术会议论文集（B 辑）(NDBC2012）. 2012.

[5] 牟乃夏, et al. ″轨迹数据挖掘城市应用研究综述.″ 地球信息科学学报 17.10 (2015): 1136-1142.

[6] 李晓旭, et al. ″Coteries 轨迹模式挖掘及个性化旅游路线推荐.″ 软件学报 29.3 (2018): 587-598.