

基于 GPS 轨迹挖掘的兴趣地点个性化推荐方法^①

张翔宇^② * * * * * 张 强^{***} 吕明琪^{****}

(* 中国科学院计算技术研究所 北京 100190)

(** 中国科学院大学 北京 100049)

(*** 浙江工业大学计算机学院 杭州 310014)

(**** 北京赛迪时代信息产业股份有限公司 北京 100048)

摘 要 随着全球定位系统(GPS)等定位设备的普及,用户可方便地记录其 GPS 轨迹,这使得自动从用户的 GPS 轨迹中发现兴趣地点(POI)(如餐厅、商场、景点)并在用户之间进行推荐成为可能。因此,本文提出了一种基于 GPS 轨迹挖掘的兴趣地点个性化推荐方法。该方法与现有主流的兴趣地点推荐平台具有以下不同:首先,现有平台假设兴趣地点是事先已知的,而该方法通过一个层次化聚类算法从用户 GPS 轨迹中自动挖掘兴趣地点。其次,现有平台的推荐模式为平台向用户推荐,因此仅考虑了用户的偏好,忽略了用户之间的社交关联对推荐效果的影响。针对此问题,该方法基于用户交叠访问行为计算用户之间的社交信任度,基于用户访问行为的相似性计算其对兴趣地点偏好的相似度,在此基础上提出了一种能够融合用户信任度和相似度的评分算法。文本基于真实 GPS 轨迹数据对提出的方法进行了评测,实验结果表明,本文所提方法的综合推荐性能明显优于简单的基于访问数量的推荐方法、仅基于用户信任度的推荐方法及仅基于偏好相似度的推荐方法。

关键词 个性化兴趣地点(POI)推荐;基于位置社会网络;全球定位系统(GPS)轨迹;时空数据挖掘

0 引 言

随着全球定位系统(global positioning system, GPS)设备的普及,用户可方便地获取其位置并记录其轨迹。基于此,许多社交网站允许用户上传、浏览和分享其记录的 GPS 轨迹数据。然而,直接分享原始 GPS 轨迹数据存在如下问题:用户感兴趣的往往不是轨迹数据本身,而是轨迹数据中所包含的丰富信息,其中最重要的即为兴趣地点(place of interest, POI)信息(如餐厅、商场、景点),但用户难以通过浏览轨迹数据直观、方便地获取这些信息,特别是在轨迹数据集不断增大的情况下。另一方面,推荐系统

是帮助用户从大量数据中快速找到自己感兴趣信息的有效手段^[1],因此本文研究旨在提出一种能够自动从用户的 GPS 轨迹中发现兴趣地点并在用户之间进行个性化推荐的方法。

实现基于 GPS 轨迹的用户到用户的个性化兴趣地点推荐系统,主要面临的挑战如下。

(1) 现有的兴趣地点推荐系统一般是平台到用户,这些系统中的待推荐兴趣地点通常是由平台定义,因此是事先已知的,如大众点评、本地生活。然而,要求用户定义待推荐兴趣地点可行性较低,因此用户到用户的推荐系统一般无法事先定义待推荐兴趣地点。

① 国家自然科学基金(U1936215)资助项目。
② 男,1977 年生,博士,高级工程师;研究方向:计算机系统结构,大数据处理,数据挖掘;联系人,E-mail: zhang_xiangyu@qq.com (收稿日期:2019-12-30)

(2) 现有的兴趣地点推荐系统获取用户偏好的方式包括显式和隐式两大类。显式偏好获取方法主要基于用户评分和协同过滤技术^[2-3]。然而,由于用户到用户的推荐系统中待推荐兴趣地点事先不可知,因此用户评分的代价较大。隐式偏好获取方法主要基于用户签到和序列挖掘技术^[4-5]。然而,用户签到仍然要以平台为主导,难以实现用户到用户的推荐。

(3) 现有的兴趣地点推荐系统大多仅考虑用户的偏好和当前上下文因素。然而,在用户到用户的推荐场景中,推荐人与被推荐人的社交关联对推荐的质量同样重要。虽然已有一些工作在兴趣地点推荐中考虑用户间的社交关联^[6-7],但这些社交关联仍然是基于用户在平台中的好友关系,难以用于用户到用户的推荐场景。

针对以上问题,本文提出了一种基于 GPS 轨迹挖掘的用户到用户的兴趣地点个性化推荐方法,该方法由 3 个步骤实现:(1) 兴趣地点挖掘。提出一种层次化聚类算法从用户的 GPS 轨迹中依次挖掘出访问点、兴趣地点和兴趣区域。(2) 关联网络构建。由于用户接近模式可在一定程度上反映其社会关联^[8],基于用户对兴趣地点访问行为的相关性(即同一时间段访问同一地点)构建基于位置的关联网络。(3) 用户偏好估计。基于用户访问行为推测用户对兴趣地点的偏好,并结合用户间偏好相似度和位置关联度预测目标用户对各兴趣地点的兴趣度评分。该方法综合考虑用户偏好和社会关联对兴趣地点进行推荐,一方面保证了推荐结果的精确度和个性化,另一方面推荐过程无需用户评分,提高了系统的可用性。

文本的主要贡献如下。

(1) 提出了一种基于交叠访问行为的用户社会关联建模与计算方法。

(2) 提出了一种融合社会关联和兴趣偏好的兴趣地点评分算法。

(3) 基于真实 GPS 轨迹数据进行了实验。

1 相关工作

兴趣地点推荐是基于位置的服务(location-

based service, LBS) 中一项非常重要的能力,不但能够提升用户的体验,还能为服务商带来商业价值。

传统的线上推荐系统最常用的技术为协同过滤(collaborative filtering)^[9],即基于用户对商品偏好的相似度做出推荐,而对用户偏好的估计方式主要包括两类,一是基于用户对商品的评分估计其偏好,二是基于用户行为估计其偏好(如用户的浏览或购买行为)。然而,传统的线上推荐模式不能适应基于 GPS 轨迹的兴趣地点推荐任务。首先,对于基于商品评分的推荐方式,待推荐商品必须是事先已知的,不符合待推荐的兴趣地点事先未知(隐含在 GPS 轨迹中)的情况,且由于用户的运动行为发生在真实世界,要求对每个访问过的兴趣地点进行评分负担过大。其次,对于基于用户行为的推荐方式,浏览、购买等线上行为数据的获取较为容易,而用户在真实世界中对兴趣地点的访问行为数据获取较为困难。

由于用户对兴趣地点的访问发生在真实世界,许多现有的兴趣地点推荐系统根据用户的当前上下文信息(如位置、时间、状态)进行推荐。例如, Park 等人^[10]基于用户当前位置、时间及天气等上下文信息向其推荐可能感兴趣的餐馆。Bellotti 等人^[11]基于机器学习方法预测用户行为(如吃饭、购物等),并开发了一个基于用户行为的兴趣地点推荐系统。Zhang 等人^[12]首先基于用户的社会特征和时空特征估计其兴趣状态(包括常规和尝鲜),然后根据用户当前兴趣状态为其有针对性地推荐餐馆。然而,此类兴趣地点推荐系统没有考虑用户偏好,推荐缺乏个性化。

为进行个性化兴趣地点推荐,早期方法采用协同过滤手段从历史兴趣地点交互数据中挖掘用户偏好。例如, Geowhiz 系统^[2]采用一种基于用户的协同过滤算法在指定位置周围推荐兴趣地点。City-Voyager 系统^[3]从历史运动轨迹中提取用户运动模式,然后采用基于项目的协同过滤方法,结合用户运动模式向其推荐商店。Yin 等人^[13]提出了一种可有效克服数据稀疏和偏好漂移问题的兴趣地点推荐算法。Hao 等人^[14]提出了一种基于表征学习的、考虑实时信息的兴趣地点推荐算法。然而,此类兴趣

地点推荐系统仍采用基于用户评分的协同过滤技术计算地点间或用户间的相似度。

为消除评分负担,许多方法从用户对兴趣地点的访问行为序列中挖掘兴趣偏好。例如, Froehlich 等人^[15]的研究表明,用户的访问行为(如访问频率、访问代价)与对兴趣地点的偏好具有正相关。蒋宗礼等人^[4]基于负采样算法生成负样本,然后采用矩阵分解算法学习隐特征向量表征用户偏好。吕慧雅^[5]采用循环神经网络(recurrent neural network, RNN)学习用户动态变化的偏好,然后结合当前上下文实现兴趣地点推荐。苏畅等人^[16]首先计算用户对地理位置的偏好,在此基础上结合用户签到相似度进行兴趣地点推荐。Xie 等人^[17]采用图嵌入技术和时间衰减模型从用户近期访问的兴趣地点中抽取其兴趣偏好,在此基础上进行推荐。然而,上述方法均只考虑用户的兴趣偏好和当前上下文等因素进

行推荐,适合平台向用户推荐的场景。在用户向用户推荐的场景中,推荐人与被推荐人的社交关联对推荐的质量同样非常重要(如用户通常偏爱朋友的推荐)^[18],因此将社会关联信息集成到推荐系统可有效改善推荐质量^[19]。有些研究在兴趣地点推荐中考虑用户间社交关系^[6,7],但这些研究中的社交关系是直接从社交网络中获取的,无法估计用户在真实世界中的社交关系,一旦脱离某个社交网络便无法实现。

2 兴趣地点挖掘

本文采用基于图 1(a)所示的用户访问模型,采用层次化的聚类方法从用户的 GPS 轨迹数据中自下而上地挖掘出访问点、兴趣地点和兴趣区域信息。

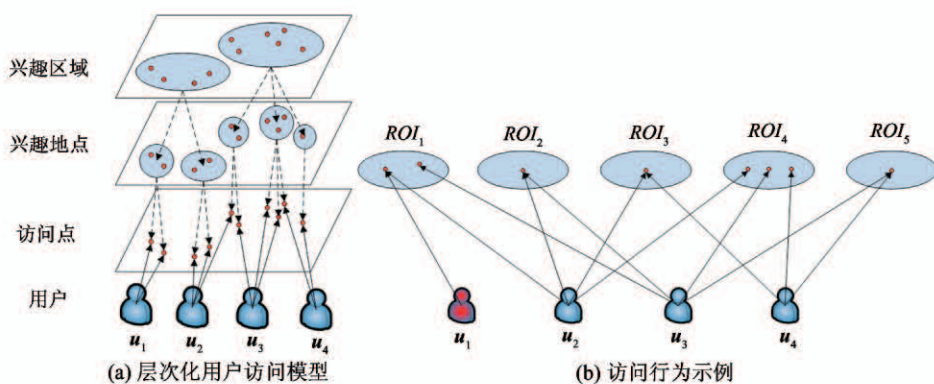


图1 用户对兴趣地点的访问模型

2.1 访问点挖掘

GPS 轨迹为一个带时间戳的 GPS 位置序列,而访问点代表用户在某一 GPS 位置周围停留较长时间(可看成是用户对某一兴趣地点的一次访问)。基于 GPS 轨迹的访问点可由两种情况产生,一是 GPS 信号丢失一段时间后又在丢失位置附近重新获取(代表室内停留),二是 GPS 位置在某一地理范围内停留一段时间(代表室外停留)。本文采用 Lv 等人^[20]提出的一种基于时间轴的聚类算法从 GPS 轨迹中统一提取这两类访问点。

2.2 兴趣地点和兴趣区域挖掘

首先,访问点代表对某个兴趣地点的一次访问,

而不同用户可能多次访问同一兴趣地点。因此,采用 DBSCAN 算法对所有访问点进行聚类,得到兴趣地点。兴趣地点为待推荐项目。其次,采用一种基于区域的聚类算法^[21]对所有兴趣地点进行聚类,得到兴趣区域。兴趣区域用于分析用户间关联,用户间关联的依据为不同用户访问同一兴趣区域。之所以不直接使用兴趣地点作为用户间关联的计算依据,是因为若兴趣地点粒度过细,则难以产生用户间关联,而若兴趣地点粒度过粗,则推荐精度会大大降低。如图 1(a)所示,许多用户基于兴趣地点无关联而基于兴趣区域则有关联(如用户 u_1 和 u_2 、 u_2 和 u_4)。因此,采用不同粒度的地理实体,分别用于生

成待推荐项目和分析用户社交关联,可有效地在用户关联覆盖和推荐精度两方面实现平衡。此外,由于 GPS 轨迹中包含家、工作地点等不适合用于推荐的私人地点,而用户往往会频繁访问其私人地点,因此首先过滤掉访问频率高于一定阈值的兴趣地点(即其包含的访问点数量大于一定阈值)。

3 基于位置的关联网络构建

3.1 关联网络构造

用户倾向于接受与其有较强社会关联用户的推荐,而这种接受的程度表现为信任度,即具有紧密社会关联的用户间通常具有较高的信任度^[22]。由于 GPS 轨迹数据不包含显式的用户关联,因此本文基于 GPS 轨迹数据中的用户访问行为相关性计算隐式的用户关联。具体来说,给定用户 u_i 和其访问点集合 VS_i ,用户 u_j 和其访问点集合 VS_j ,找出其中所有满足如下条件的访问点对 V_i 和 V_j :(1) $V_i \in VS_i$, $V_j \in VS_j$;(2) V_i 和 V_j 属于同一兴趣地点;(3) V_i 和 V_j 在时间上存在交叠。

如果至少存在两对以上的访问点满足上述条件,则认为用户 u_i 和用户 u_j 存在社交关联(之所以要求存在两对以上,是为了避免无真实社会关系的用户偶然同时访问同一兴趣地点所产生的关联噪声),而用户 u_i 和用户 u_j 的社交关联权值 w_{ij} 为所有满足上述条件的访问点对的交叠时间的总和。在计算得到所有用户之间的社交关联之后,可以用如下的基于位置的关联网络进行表示。

定义 1(基于位置的关联网络) 基于位置的关联网络为一个带权图 $G=(V, E)$,顶点集合 $V=\{u\}$ 代表用户,边集合 $E=\{e\}$ 代表用户关联,边 $e_{ij}=(u_i, u_j, w_{ij})$ 中的 u_i, u_j 为相关联的两个用户, w_{ij} 为两个用户间的关联度权值。

根据定义 1,用户间基于位置的关联是对称的,即用户 u_i 和 u_j 的关联与 u_j 和 u_i 的关联相同,则基于位置的关联网络可由一个对称的邻接矩阵表示。以图 1(b)为例,兴趣区域 ROI_k 中的圆点代表兴趣地点,从用户指向兴趣地点的线段代表访问行为。若两个用户的访问线段指向同一兴趣地点则代表访

问时间存在交叠。假设每次交叠访问的时间均为 1,则图 1(b)所示的用户访问行为可得到式(1)的邻接矩阵 N 表示的基于位置的关联网络,其中 $N[i][j]$ 为 u_i 和 u_j 的关联度权值(为便于展示,本例没有附加要求至少存在两对以上交叠访问点的限制)。

$$N = \begin{matrix} & \begin{matrix} u_1 & u_2 & u_3 & u_4 \end{matrix} \\ \begin{matrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{matrix} & \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \end{matrix} \tag{1}$$

3.2 信任度计算

关联网络中关联用户间的信任度可由其关联度权值表示。然而,随着网络规模的增大,能直接产生交叠访问行为的用户比例将变得很小,用户间的直接关联将变得极度稀疏,而关联度权值只存在于有直接关联的用户之间,导致网络中大部分用户间的信任度是缺失的。为解决这一问题,本文通过扩展 PageRank 算法提出了一种信任度传播算法,用于计算关联网络中没有直接关联的用户间的信任度。该算法的思想是:(1) 对于关联网络中的用户 u_i 和 u_j ,只要 u_i 和 u_j 之间存在可达的路径,则 u_i 和 u_j 之间存在关联;(2) u_i 和 u_j 之间的信任度由路径中所有用户的关联度权值进行随机游走传播得到。例如,用户 u_1 和 u_2 有直接关联,用户 u_2 和 u_3 有直接关联,则用户 u_1 和 u_3 有间接关联。图 1(b)中用户 u_1 和 u_3 也可通过中间用户 u_2 产生间接关联。

标准 PageRank 算法不应用户信任度计算任务,这是由于标准 PageRank 算法最初用于计算网页的重要度,它对整个网络统一地进行信任度传播,缺乏个性化(例如,标准 PageRank 算法计算得到的网页重要度对任何用户都一样),这显然不符合个性化推荐的要求。针对此问题,本文对标准 PageRank 算法进行扩展,使其针对不同的目标用户采取不同的随机游走策略,具体如下:(1) 算法强制以目标用户为起点开始随机游走,其目的是使得与目标用户有直接关联的其他用户拥有最高的初始信任度。(2) 某个用户的信任度并不平均传播到与其相关联的所有用户,而是按照关联度权值的比例进行传播,

其目的是使得关联度权值高的其他用户获得更高的信任度。

该信任度传播算法如算法 1 所示,其中 w_{ij} 为用户 u_i 和 u_j 的关联度权值。向量 tv 和 ctv 分别用于存储目标用户对其他用户上次和本次迭代中的信任度,其中 tv 和 ctv 的第 1 个元素用于存储目标用户对本身的信任度。算法从目标用户开始,把其所拥有的信任度按关联度权值比例不断传播给直接关联的用户(第 5~9 行),直到算法收敛。其中,参数 α 的值若设置的较小则与目标用户关联较疏远的用户(两个用户在关联网络中的路径距离较大)将获得更低的信任度,且算法收敛速度更快,但同时会减少与目标用户产生信任度的用户数量,导致实际推荐时所能考虑的信息变少。本文将 α 的值设置为 0.5。该算法的时间复杂度为 $O(t(\alpha) |V|^2)$, 其中 $|V|$ 为关联网络中节点的数量, $t(\alpha)$ 为算法的迭代次数(与参数 α 有关)。

算法 1 信任度传播算法

输入: 基于位置的关联网络 G , 阻尼系数 $\alpha (0 < \alpha < 1)$

输出: 信任度向量 tv ($tv[i]$ 代表目标用户对用户 i 的信任度)

```

1: last trust scores vector  $tv$ , current trust vector  $ctv$ 
2:  $n = \text{size of } G.V$ ,  $tv = 0^n$ ,  $tv[0] = 1$ 
3: while algorithm has not converged do
4:    $ctv = 0^n$ ,  $ctv[0] = 1 - \alpha$ 
5:   for each  $i \in G.V$  do
6:      $out\_weight = \sum_{j \in G.V} w_{ij}$ 
7:     if  $out\_weight > 0$  then
8:       for each  $j$  such that  $e_{ij} \in G.E$  do
9:          $ctv[j] = ctv[j] + \frac{\alpha \times w_{ij} \times tv[i]}{out\_weight}$ 
10:   $tv = ctv$ 
```

4 兴趣地点兴趣度评分计算

4.1 兴趣地点偏好度计算

用户对兴趣地点的偏好是实施推荐的最本质因素。因此,为实现有效的用户间推荐,还必须考虑不同用户对兴趣地点偏好的相似度。Froehlich 等人^[15]

的研究发现,用户对某个兴趣地点的访问次数与其对该兴趣地点的偏好具有正相关,而用户对某个兴趣地点的访问代价与其对该兴趣地点的偏好也具有正相关(即用户愿意花费大的代价去访问某个兴趣地点,说明其对该兴趣地点非常偏好)。因此,为从 GPS 轨迹中获取用户对兴趣地点的偏好,可考虑其对兴趣地点的访问次数和访问代价。由于访问代价难以获取,本文仅基于访问次数估计用户对兴趣地点的偏好。然而,由于兴趣地点粒度较小,导致不同用户访问同一兴趣地点的情况难以产生。针对此问题,可首先用兴趣区域代替兴趣地点,并考查不同用户访问同一兴趣区域的情况。

然后,基于用户对兴趣区域和对兴趣区域中兴趣地点的访问次数,计算用户对每个兴趣地点的偏好度。给定目标用户,其对兴趣区域 i 中的兴趣地点 j 的偏好度 PS_{ij}^i 由式(2)计算得到。其中, T_u 为目标用户对用户 u 的信任度, V_u^i 为用户 u 对兴趣区域 i 的访问次数, $V_u^{i,j}$ 为用户 u 对兴趣区域 i 中的兴趣地点 j 的访问次数, RS_i 为目标用户对兴趣区域 i 的偏好度。式(2)的思想为将目标用户对兴趣区域 i 的偏好度加权分配给其中包含的兴趣地点,而目标用户对兴趣区域 i 的偏好度 RS_i 的计算方法见 4.2 节。

$$PS_{ij}^i = \frac{\sum_{u \in G.V} T_u \times V_u^{i,j}}{\sum_{u \in G.V} T_u \times V_u^i} \times RS_i \quad (2)$$

4.2 兴趣区域兴趣度计算

为计算兴趣区域的偏好度,最简单的方法是直接采用用户对兴趣地点的访问次数。然而,这种方法存在一个问题,就是系统只能计算用户对其访问过的兴趣地点的偏好度,无法估计用户对其未访问过的兴趣地点的偏好度。针对此问题,本文提出一种基于 HITS^[23] 的偏好度计算算法,其主要思想为: (1) 与目标用户具有类似偏好的其他用户的偏好可用于隐式地估计目标用户的偏好(与协同过滤的思想类似)。(2) 假定用户与兴趣区域间存在互增强关系(访问更重要兴趣区域的用户更重要,而被更重要用户访问的兴趣区域更重要,更重要的用户或兴趣区域对目标用户偏好的影响更大)。与 HITS 算法类似,用户被看作是 hub,兴趣区域被看作是

authority, 而 hub 值较高的用户访问的兴趣区域有较高的 authority 值, 而访问具有较高 authority 值的用户通常 hub 值较高。HITS 算法包含两个主要步骤, forward 步骤, 一个兴趣区域获得的 authority 值与所有访问它的用户的 hub 值之和成正比; backward 步骤, 一个用户获得的 hub 值与所有他访问的兴趣区域的 authority 之和成正比。

然而, 标准 HITS 算法不能适应兴趣区域偏好度计算任务, 这是由于, 一方面, 标准 HITS 算法以统一的方式进行 hub 值和 authority 值的分配, 计算结果对所有用户均相同, 缺乏个性化, 而实际上不同用户对同一兴趣区域的偏好度是不同的; 另一方面, 标准 HITS 算法只考虑用户和兴趣区域间的关联 (即用户对兴趣区域的偏好), 忽略用户之间的关联 (即用户之间的信任度), 而本文的研究目的在于融合用户间的信任度和偏好相似度进行推荐。针对这些问题, 本文对标准 HITS 算法做如下扩展。

(1) 为解决标准 HITS 算法缺乏个性化的问题, 在算法初始化时, 将除指定用户外所有用户的 hub 值初始化为 0, 而指定用户的 hub 值初始化为 1。这样, 在第 1 次 forward 步骤中指定用户偏好的兴趣区域将获得较多的 authority 值, 从而与指定用户偏好相似的用户也将获得较多的 hub 值。此外, 在 backward 步骤中采用一个阻尼系数 β 来限制所有非指定用户所获得的 hub 值。

(2) 为将用户间信任度集成到兴趣度评分算法中, 算法在 backward 步骤中将所有用户获得的 hub 值乘上其信任度, 则信任度高的用户将获得更多的 hub 值, 从而其偏好的兴趣区域也将获得更多的 authority 值。

扩展后的兴趣度评分算法如算法 2 所示, 其中 TV 是由信任度传播算法计算得到的所有用户的信任度权值向量, 用户访问行为图 $G = (U, R, E)$, 顶点集合 U 代表用户, 顶点集合 R 代表兴趣区域, 边集合 E 代表用户对兴趣区域的访问行为, 边 $e_{ij} = (u_i, r_j, w_{ij})$ 代表用户 u_i 对兴趣区域 r_j 的访问次数是 w_{ij} 。向量 hv 和 av 分别用于存储所有用户的 hub 值和所有兴趣区域的 authority 值, 而 hv 第 1 个元素用于存储指定用户的 hub 值。算法反复进行 for-

ward 步骤 (第 4 ~ 6 行) 和 backward 步骤 (第 7 ~ 10 行) 直到收敛。与参数 α 类似, 参数 β 设置的较小则算法的个性化程度增强 (即指定用户将获得更高的 hub 值, 而其访问的兴趣区域将获得更高的 authority 值), 但多样化程度降低 (即用户未访问过的兴趣区域的 authority 值急剧降低, 导致某些兴趣区域无法得到推荐)。本文将 β 的值设置为 0.8。该算法的时间复杂度为 $O(t(\beta) |U| |R|)$, 其中 $|U|$ 为访问行为图中用户的数量, $|R|$ 为访问行为图中兴趣区域的数量, $t(\beta)$ 为算法的迭代次数 (与参数 β 有关)。

算法 2 兴趣区域偏好度计算算法

输入: 信任度向量 TV , 用户访问行为图 G , 阻尼系数 $\beta (1 < \beta < 1)$

输出: 兴趣度向量 av ($av[i]$ 代表指定用户对兴趣区域 i 的兴趣度)

```

1: hub score vector  $hv$ , authority score vector  $av$ 
2:  $n = \text{size of } G.U, m = \text{size of } G.R, hv = 0^n, hv[0] = 1$ 
3: while algorithm has not converged do
4:    $av = 0^m$  // forward step
4:   for each  $i \in G.R$  do
5:     for each  $u \in G.U$  do
6:        $av[i] = av[i] + \frac{w_{ui} \times hv[u]}{\sum_{j \in G.R} w_{uj}}$ 
7:    $hv = 0^n, hv[0] = 1 - \beta$  // backward step
8:   for each  $u \in G.U$  do
9:     for each  $i \in G.R$  do
10:       $hv[u] = hv[u] + \frac{\beta \times w_{ui} \times TV[u] \times av[i]}{\sum_{v \in G.U} w_{vi} \times TV[v]}$ 

```

5 实验

5.1 实验设置

采集本文基于真实用户 GPS 轨迹数据对提出的推荐方法进行评测。实验采用 Android 智能手机作为数据采集设备, 将 GPS 轨迹数据采集程序运行在智能手机上。通过大约 4 个月的时间, 采集了 14 位实验参与者的真实 GPS 轨迹, 其中部分参与者为互相熟悉的同学或同事。参与者每访问一个地点, 都通过自报告的方式使用采集程序记录下该地点的信息 (如名称、类型等)。最终, 数据集包含 2416 个

访问点记录(实验参与者每天平均记录的访问点个数为 1.4),实验参与者每天平均的行走距离为 5.3 km。

在兴趣地点挖掘过程中,一方面考虑到兴趣地点形状的任意性,实验采用基于密度的聚类算法 DBSCAN 从访问点集中提取兴趣地点,另一方面为限制兴趣区域的大小,实验采用基于区域的聚类算法^[21]从兴趣地点集中提取兴趣区域,并将兴趣区域限制在边长为 500 m 的正方形内。实验最终得到 2566 个访问点、497 个兴趣地点和 181 个兴趣区域。

实验采用用户问卷调查的方法评测推荐系统的有效性。对每一个参与者,系统基于提出的算法从得到的所有兴趣地点中提取出该参与者偏好度最高的 10 个兴趣地点,并基于这些兴趣地点要求该参与者完成相应的评价调查。评价调查要求参与者对得到的 10 个兴趣地点进行打分和排序,其中对每个兴趣地点的打分标准为,1 分对应“对此兴趣地点感兴趣”,0 分对应“不熟悉此兴趣地点”,-1 分对应“对此兴趣地点不感兴趣”。

5.2 推荐性能评测

实验评测推荐系统寻找兴趣地点及对兴趣地点排序的能力。一方面,基于用户打分结果计算所获取 10 个兴趣地点的准确率(Precision)值,用以评测推荐系统寻找有效兴趣地点的能力。其中,有效兴趣地点即分值为 1 的兴趣地点。Precision 即为所获取的 10 个兴趣地点中有效兴趣地点所占的比率。另一方面,基于用户对兴趣地点的排序结果计算所获取 10 个兴趣地点的 nDCG(normalized discounted cumulative gain)值,用以评测推荐系统对兴趣地点进行排序的能力。nDCG 的计算公式如式(3)所示,其中 p 为待排序的兴趣地点数量(即 10), rel_i 是排在第 i 位的兴趣地点的相关值。在本实验中,某个兴趣地点相关值为: p -用户对其排序序号+1(如用户排序序号为 2 的兴趣地点相关值为 9)。而 $IDCG[p]$ 为在用户排序情况下(即理想情况下)的 $DCG[p]$ 值。

$$nDCG[p] = \frac{DCG[p]}{IDCG[p]}$$

$$DCG[p] = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i} \quad (3)$$

实验计算 14 位参与者的平均 Precision 与 nDCG[10]值,并与简单的基于访问数量的推荐方法 Rank by Count(即优先推荐被访问次数多的兴趣地点)进行比较,结果如表 1 所示。从表中可以看出本文提出的推荐方法具有较高的 Precision 和 nDCG 值,寻找兴趣地点和对兴趣地点排序的能力明显优于基于访问数量的推荐方法。

表 1 推荐系统的性能评测

	Precision	nDCG[10]
本文方法	72.9%	0.954
Rank by Count	48.6%	0.782

5.3 比较实验

为评测本文提出的推荐方法的有效性,将其与如下方法进行比较。

(1) Rank by Preference。通过将兴趣度评分算法中用户信任度强制设置为相同,仅考虑用户间偏好相似度进行兴趣度计算。

(2) Rank by Trust。仅依赖信任度传播算法计算用户间信任度,并优先推荐信任度高用户偏好的兴趣地点。

(3) Rank by CF。基于协同过滤的方法^[4]。首先,基于用户对兴趣地点的访问行为构建用户对兴趣地点的访问行为矩阵 UPM ,其中 $UPM[i][j]$ 代表用户 u_i 对兴趣地点 p_j 的访问次数。然后,采用矩阵分解算法学习用户和兴趣地点的隐特征向量,在此基础上通过对 UPM 进行协同过滤实现推荐。

(4) Rank by GE。基于图嵌入的方法^[17]。首先,构造用户对兴趣地点的访问行为图 UPG ,其中用户 u_i 与兴趣地点 p_j 之间的边代表 u_i 对 p_j 的访问(权重为访问次数)。然后,基于图嵌入算法计算所有用户和兴趣地点的隐向量表示。最后,基于隐向量点乘计算推荐分数。

实验比较结果如表 2 所示。从表中可以看出,首先,本文提出的方法的推荐准确率高于一因素推荐方法(即仅基于用户偏好的 Rank by Preference 和仅基于社交关联的 Rank by Trust),而对兴趣地点的排序能力明显优于 Rank by Trust。其

次,本文提出的方法在推荐性能上与 Rank by CF 和 Rank by GE 相差不大,这说明在单纯兴趣地点推荐的场景下(一般指平台到用户的推荐场景),用户偏好是影响推荐质量的主要因素,只要能较好地挖掘和表征用户偏好,即能获得理想的推荐性能。

表 2 不同推荐方法的推荐性能比较

	Precision	nDCG[10]
本文方法	72.9%	0.954
Rank by Preference	61.4%	0.939
Rank by Trust	64.3%	0.784
Rank by CF	71.8%	0.945
Rank by GE	72.5%	0.951

以上实验中,参与者仅知道兴趣地点本身信息,而不知道该兴趣地点被哪些其他参与者访问过,为考查信任度对兴趣地点推荐结果接受程度的影响,参与者在被告知兴趣地点被哪些其他参与者访问过的情况下进行第 2 次评价,结果如表 3 所示(其中 1st precision 和 2nd precision 分别代表第 1 次和第 2 次评价结果的 Precision 值)。从表中可以看出:首先,本文提出的方法与 Rank by Trust 第 2 次评价结果的 Precision 值明显高于第 1 次,这是由于一些参与者在知道兴趣地点被访问信息的情况下对原先自己不熟悉(即分值为 0)但被其好友访问过的兴趣地点开始感兴趣。而 Rank by Preference 不考虑用户社会关系,所以 Precision 值提高不明显。其次,本文提出的方法在该场景下与 Rank by CF 和 Rank by GE 的性能差距拉大,这是由于 Rank by CF 和 Rank by GE 也是只考虑用户偏好,没有考虑社交关联在用户到用户的推荐场景中对推荐接受度的潜在影响。

表 3 访问者信息透明情况下不同推荐方法的推荐性能比较

	1st precision	2nd precision
本文方法	72.9%	84.3%
Rank by Preference	61.4%	62.9%
Rank by Trust	64.3%	74.3%
Rank by CF	70.8%	74.9%
Rank by GE	72.1%	75.4%

6 结 论

针对现有兴趣地点推荐方法只能实现平台到用户推荐模式的不足,本文提出了一种基于 GPS 轨迹挖掘的兴趣地点个性化推荐方法,可实现用户到用户推荐模式。为实现用户到用户推荐模式,该方法一方面通过层次化聚类算法自动从用户 GPS 轨迹中挖掘兴趣地点,另一方面提出了一种能够融合用户之间社交信任度和兴趣相似度的推荐评分算法。基于真实数据的实验结果表明,在用户已知推荐者的场景下,该方法的性能明显优于现有方法。

参考文献

[1] Lu J, Wu D, Mao M, et al. Recommender system application developments: a survey[J]. *Decision Support Systems*, 2015, 74: 12-32

[2] Horozov T, Narasimhan N, Vasudevan V. Using location for personalized POI recommendations in mobile environments[C] // *Proceedings of SAINT*, Washington, USA, 2006: 124-129

[3] Takeuchi Y, Sugimoto M. CityVoyager: an outdoor recommendation system based on user location history [C] // *Proceedings of UbiComp*, Orange, USA, 2006: 625-636

[4] 蒋宗礼, 王逸鹤. 基于序列挖掘的兴趣点推荐算法[J]. *北京工业大学学报*, 2019, 45(9): 853-858

[5] 吕慧雅. 基于循环神经网络的兴趣点推荐模型设计与实现[D]. 广州:华南理工大学计算机学院, 2018

[6] 温彦, 马立健, 曾庆田, 等. 基于地理信息偏好修正和社交关系偏好隐式分析的 POI 推荐[J]. *数据分析与知识发现*, 2019, 3(8): 30-40

[7] 李昱杭, 杨艳, 高静远. 地理分层结构与社会关系的兴趣点推荐[J]. *软件工程*, 2019, 22(10): 12-18

[8] Terry M, Mynatt E, Ryall K, et al. Social Net: using patterns of physical proximity over time to infer shared interests [C] // *Proceedings of CHI*, Minnesota, USA, 2002: 816-817

[9] Herlocker J, Konstan J, Terveen L, et al. Evaluating collaborative filtering recommender systems[J]. *ACM Transactions on Information Systems*, 2004, 22(1): 5-53

[10] Park M, Hong J, Cho S. Location-based recommendation system using Bayesian user's preference model in mobile devices[C] // *Proceeding of UIC*, Hong Kong, China, 2007: 1130-1139

[11] Bellotti V, Begole B, Chi E, et al. Activity-based serendipitous recommendations with the Magitti mobile leisure

- guide[C]//Proceedings of CHI, Florence, Italy, 2008: 1157-1166
- [12] Zhang F, Zheng K, Yuan N, et al. A novelty-seeking based dining recommender system[C]//Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 2015: 1362-1372
- [13] Yin H, Wang W, Wang H, et al. Spatial-aware hierarchical collaborative deep learning for POI recommendation[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(11): 2537-2551
- [14] Hao P, Cheang W, Chiang J. Real-time event embedding for POI recommendation[J]. *Neurocomputing*, 2019, 349(15): 1-11
- [15] Froehlich J, Chen M, Smith I, et al. Voting with your feet: an investigative study of the relationship between place visit behavior and preference[C]//Proceedings of UbiComp, Orange, USA, 2006: 333-350
- [16] 苏畅, 武鹏飞, 谢显中, 等. 基于用户兴趣和地理因素的兴趣点推荐方法[J]. *计算机科学*, 2019, 46(4): 228-234
- [17] Xie M, Yin H, Wang H, et al. Learning graph-based POI embedding for location-based recommendation[C]//Proceedings of CIKM, Indianapolis, USA, 2016: 15-24
- [18] Draidi F, Pacitti E, Cart M, et al. Leveraging social and content-based recommendation in P2P systems[C]//Proceedings of AP2PS, Lisbon, Portugal, 2011: 13-18
- [19] Yuan T, Cheng J, Zhang X, et al. How friends affect user behaviors? an exploration of social relation analysis for recommendation[J]. *Knowledge-Based Systems*, 2015, 88: 70-84
- [20] Lv M, Chen L, Xu Z, et al. The discovery of personally semantic places based on trajectory data mining[J]. *Neurocomputing*, 2016, 173: 1142-1153
- [21] Zheng Y, Xie X, Yang Q. Collaborative location and activity recommendations with GPS history data[C]//Proceedings of the 19th International Conference on World Wide Web, Raleigh, USA, 2010: 1029-1038
- [22] Ziegler C, Georg L. Analyzing correlation between trust and user similarity in online communities[C]//Proceedings of International Conference on Trust Management, Oxford, UK, 2004: 251-265
- [23] Kleinberg J. Authoritative sources in a hyperlinked environment[J]. *Journal of the ACM*, 1999, 46(5): 604-632

Personalized POI recommendation based on GPS trajectory mining

Zhang Xiangyu^{* ** ****}, Zhang Qiang^{* **}, Lv Mingqi^{***}

(^{*} Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

(^{**} University of Chinese Academy of Sciences, Beijing 100049)

(^{***} College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310014)

(^{****} Beijing CCID Info Tech. Inc, Beijing 100048)

Abstract

With the increasing availability of global positioning system (GPS) enabled devices, users can easily record their GPS trajectories. It opens the possibility of discovering place of interests (POIs) (e.g., restaurants, malls, scenic spots) from GPS trajectories and recommending between users. Therefore, a method for personalized POI recommendation based on GPS trajectory mining is proposed. The proposed method differs from existing POI recommendation platforms in the following aspects. First, the existing platforms assume that the POIs are known in advance, while the proposed method automatically extracts POIs from GPS trajectories based on a hierarchical clustering algorithm. Second, the existing platforms only consider the users' preference, ignoring the effect of social relations between users on the recommendation quality. Aiming at this problem, the proposed method estimates the social trust degree based on overlapping visiting activities and estimates the preference similarity on POIs based on the similarity of visiting activities. Then, it proposes an algorithm to fuse the social trust degree and POI preference similarity. The proposed method is evaluated based on real GPS trajectory dataset. The experiment results show that the overall performance of the proposed method outperforms the simple recommender based on visit counts, the recommender based only on social trust degree, and the recommender based purely on POI preference similarity.

Key words: personalized place of interest (POI) recommendation, location-based social network, global positioning system (GPS) trajectory, spatiotemporal data mining