LBSNs中的群体行程推荐方法

李效伦 丁志军
(同济大学嵌入式系统与服务计算教育部重点实验室 上海 201804)

摘 要 随着GPS设备(如智能手机、GPS导航仪、GPS记录仪等)的广泛应用,其产生的位置信息也越来越多。基于位置的社交网络(Location-Based Social Networks,LBSNs)推荐系统受到了更多的关注。旅游行程推荐是LBSNs中非常热门的研究课题之一,但是现有研究主要侧重向单个用户推荐旅游行程,缺乏向群体推荐行程的工作。因此提出了一种LBSNs中的群体行程推荐方法。该方法首先根据用户的签到记录,使用K-means和谱聚类方法挖掘用户群体及其偏好;然后综合考虑群体对行程的时间和价格的约束,设计了行程推荐算法向群体用户推荐符合其偏好的旅游行程;最后,使用新浪微博用户的真实签到记录进行实验分析,结果表明所提出的群体行程推荐方法具有良好效果。

关键词 群体,旅游行程,LBSNs,谱聚类,K-means,推荐系统
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.06.033

Group Travel Trip Recommendation Method in LBSNs

LI Xiao-lun DING Zhi-jun
(The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China)

Abstract With the widespread adoption of GPS-enabled devices, such as smartphone, GPS navigation device, GPS logger, etc., more and more location information is collected. Recommender systems for location-based social networks (LBSNs) have received more attention. The research on recommending a trip to a group is a hot topic, but most related works mainly focus on recommending trip to a user and lack in recommending trip to a group. Therefore, this paper proposed a trip recommender method for a group in LBSNs. First, according to user's check-ins, the proposed method uses K-means and spectral clustering to mine groups who have a great many same check-ins. Then, group's preference is obtained based on their common check-ins. At last, combining group's time and cost constraint, a trip recommender algorithm is designed to recommend trip which satisfies group's preference to a group. This paper conducted experiments with users' real check-ins of Sina weibo. The experimental results show that the proposed method in this paper to recommend trips to a group achieves good effects.

Keywords Group, Travel trip, LBSNs, Spectral clustering, K-means, Recommender system

1 引言

移动通信技术(如3G,4G,WiFi等)的快速发展以及移动终端设备(如智能手机、PDA等)的普及,使得越来越多的移动用户在任何时候、任何地点都可以访问互联网。在用户移动过程中,可以通过GPS、基站或者WiFi定位技术获取用户的地理位置信息或移动轨迹。大量的位置信息和移动轨迹推动了整个LBSNs推荐系统的快速发展。LBSNs是指基于位置的社交网络,例如Foursquare、新浪微博等;同时,LBSNs推荐系统也给人们带来了更多的便利,例如附近停车场、饭店、加油站、旅游行程等的推荐。近年来,国内外已经有许多研究人员在LBSNs的位置推荐[1-4]、活动推荐[4,6-7]和行程推荐[5,8,13-14]方面展开了深入的研究。其中,向单个用户推荐旅游行程是非常热门的研究方向之一。Zhang等人[8]提出一个推荐旅游路线的原型系统,该系统用于推荐从一个旅游景点到另一个旅游景点的路线。在推荐过程中,其不仅考虑路线长短和交通问题,同时还考虑路线上有没有美丽的风景。Kurashima等人[13]首先使用均值漂移算法[9]从Flickr用户上传的带有地理标签(时间、位置等)的图片中提取地标,然后使用马尔可夫模型[10]和主题模型[11-12]构建用户的概率行为模型;在此基础上,提出了一个向单个用户推荐旅游行程的方法。Lu等人[14]综合考虑用户的偏好、时间和价格约束,使用协同过滤的方法向单个用户推荐旅游行程。目前,群体发现已经有不少的研究。例如,Wang等人[15]提出基于群体的个性化位置推荐方法,首先根据用户的历史位置构建其加权分类层次结构(WCH),然后根据其WCH使用K-means算法将

到稿日期:2016-05-31 返修日期:2016-09-21 本文受国家自然科学基金项目(61173042,61472004)资助。
李效伦(1990—),男,硕士,主要研究方向为基于位置的社交网络推荐系统,E-mail:lxl910915@163.com;丁志军(1974—),男,博士,教授,主要研究方向为Petri网、服务计算。

用户划分到不同的群体，构造用户群体。He 等人[16]提出了一种 GPS 轨迹划分方法，使用基于频率的方法根据用户的历史 GPS 轨迹挖掘其频繁走过的路线，将具有相似路线的用户定义为一个群体。Lin 等人[17]提出元图分解框架，从各种各样的社交内容和社交关系中提取社区结构。作者提出了元图的概念，元图是一个关系型超图，它表示以不同的方式连接多种不同的对象的实体（如社交网络中的标签、评论、照片、视频、文本），然后从元图中发现潜在的社区结构。但是，在 LBSNs 领域对群体发现的研究还存在不足。

综上所述，现有研究主要侧重于面向单个用户推荐旅游行程，向群体用户推荐旅游行程的研究相对缺乏。事实上，在有效挖掘群体及其偏好的基础上向群体用户推荐符合群体偏好的旅游行程，可以为旅行社定制旅行团旅游方案、用户组团个性化旅游等提供技术支撑。因此，开展向群体推荐旅游行程的方法研究是十分必要的。本文提出了一种 LBSNs 中的群体行程推荐方法。该方法首先根据用户的签到记录，使用 K-means 和谱聚类方法挖掘用户群体及其偏好；然后综合考虑群体行程的时间和价格约束，设计了行程推荐算法向群体用户推荐符合其偏好的旅游行程。本文主要工作有以下 4 点：

1）给出了向群体推荐行程的技术框架，该框架包含 3 个部分，即群体发现、群体偏好获取以及行程推荐；

2）根据用户的签到记录分别使用 K-means 和谱聚类方法挖掘群体，并基于群体中用户在不同分类的景点上的签到数据获取群体的偏好；

3）给出了基于群体偏好的群体行程推荐算法，该算法根据群体的偏好、当前位置、群体的时间约束和群体的价格约束向其推荐旅游行程；

4）使用新浪微博用户的实际签到数据进行实验分析，结果表明本文所提方法具有良好的向群体推荐行程的效果。

本文第 2 节给出一些相关的定义和符号；第 3 节给出使用 K-means 和谱聚类方法挖掘群体的过程；第 4 节介绍向群体推荐行程的算法，并通过一个示例阐述推荐过程；第 5 节介绍数据集的获取，并通过真实的数据集验证本文提出的框架的推荐效果，通过不同的对照实验分析不同方法的实验结果；最后对本文工作进行总结和展望。

2 相关定义

本节主要声明了文中使用的一些定义和符号。

定义 1（景点） L 表示所有景点的集合，景点信息主要包含：景点名称，经纬度，分类 c_l，得分 S_l，价格 C_l，游玩时间 T_l 等。

图 1 示出 5 个景点 l_1,l_2,l_3,l_4,l_5 以及群体 U 的位置分布。景点间的实线表示从一个景点到另外一个景点所花费的时间 T，该时间可以通过两个景点的经纬度进行换算（或者通过谷歌地图、高德地图提供的 API 获取），例如 $T_{l_1l_5}=1$，表示从 l_1 到 l_5 的时间为 1 小时。虚线表示群体 U 到各个景点所需要的时间，例如 $T_{Ul_1}=1$。表 1 列出了 5 个景点 l_1,l_2,l_3,l_4,l_5 的分类、游玩时间、价格以及得分。

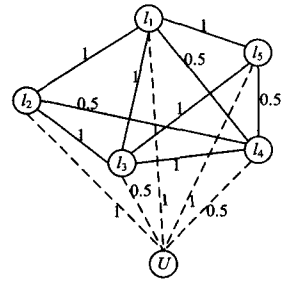

图 1 景点地图示例

表 1 景点信息

景点	分类	游玩时间	价格	得分
l_1	c_1	2	10	4.5
l_2	c_2	3	90	2.0
l_3	c_3	3	70	3.5
l_4	c_4	2	120	4.0
l_5	c_3	1	100	4.0

定义 2（行程） $U\rightarrow l_1\rightarrow l_2\rightarrow\cdots\rightarrow l_m\rightarrow U$ 表示一个行程，行程中包含多个景点。

例如，行程 $U\rightarrow l_1\rightarrow l_5\rightarrow U$ 表示群体 U 首先访问了景点 l_1，然后访问了景点 l_5，最后回到了 U 的出发点。

定义 3（用户） u 表示一个用户。

图 2 中的用户-位置两层图表示用户和位置间的关系，用户层中有 4 个用户 u_1,u_2,u_3,u_4；位置层中有 5 个景点 l_1,l_2,l_3,l_4,l_5。用户层到位置层的有向箭头上的值表示用户在相应景点上的签到总次数。

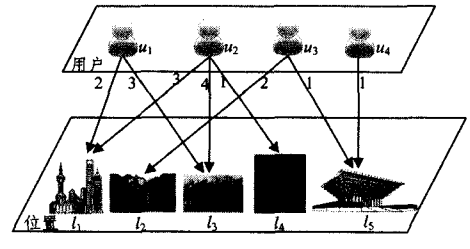

图 2 用户-位置两层图

定义 4（用户签到的景点） L_u 表示用户 u 签到过的景点的集合，$L_u=\{l_1,l_2,\cdots,l_k\}$。

例如，在图 2 中，u_1 的签到历史为 $L_{u_1}=\{l_1,l_3\}$。

定义 5（用户签到向量） 使用一个 n 维向量 $I_u=(I_{u,1},I_{u,2},\cdots,I_{u,|L|})$ 表示用户 u 在所有景点 L 上的签到向量，向量中第 i 个元素的值表示用户在第 i 个景点上的签到次数。如果第 i 个元素的值为 0，则表示用户 u 没有在第 i 个景点上签到。在图 2 中，$I_{u_1}=(2,0,3,0,0)$。为了避免一些用户在一个位置的签到次数过多而影响聚类结果，对用户签到向量进行归一化处理。用户签到向量中每一位置的值都除以该位置上签到次数最多的用户的签到次数，表示为归一化签到向量 $\hat{I}_u$。

定义 6（用户相似度） 用户 u 和用户 v 的相似度为两个用户的归一化签到向量的余弦相似度，可以使用式(1)计算得到。

$$\hat{W}_{u,v}=\frac{\sum_{l\in L}\hat{I}_{u,l}\hat{I}_{v,l}}{\sqrt{\sum_{l\in L}\hat{I}^2_{u,l}}\sqrt{\sum_{l\in L}\hat{I}^2_{v,l}}} \tag{1}$$

定义 7(群体)　$U=\{u_1,u_2,\cdots,u_k\}$表示一个群体，群体中包含了多个用户。

这些用户共同签到的景点比较多。即，如果多个用户共同签到的景点较多，本文则倾向于将这些用户定义为一个群体。

定义 8(群体偏好)　x 维向量 $P_U=(p_1,p_2,\cdots,p_x)$表示群体 U 的偏好。x 表示景点分类的个数，向量 P_U 中的第 j 个元素的值表示群体对第 j 个景点分类的偏好程度，元素的取值范围为[0,1]。根据群体中用户共同签到的景点的分类，将每个分类所占的比例作为群体对该分类的偏好。对于群体 $U=\{u_1,u_2\}$，$P_U=\{0.5,0,0.5,0\}$。

定义 9(时间和价格约束)　群体 U 设定的游玩时间以及可以接受的最高价格。

以图 1 为例，可以设定群体 U 的时间约束为 8，价格约束为 200。

定义 10(行程的时间)　行程 $U\rightarrow l_1\rightarrow l_2\rightarrow\cdots\rightarrow l_m\rightarrow U$ 的时间包括从群体 U 到 l_1 的时间、从 l_1 依次到 l_m 的时间、在 l_1 到 l_m 的游玩时间以及从 l_m 回到群体 U 的起始位置的时间：

$$T_{Ul_1}+\sum_{i=1}^{m-1}T_{l_il_{i+1}}+\sum_{i=1}^{m}T_{l_i}+T_{l_mU} \tag{2}$$

其中，T_{l_i} 表示景点 l_i 的游玩时间。

在图 1 中，行程 $U\rightarrow l_1\rightarrow l_5\rightarrow U$ 的时间为 1+1+2+1+1=7。

定义 11(行程的价格)　行程 $U\rightarrow l_1\rightarrow l_2\rightarrow\cdots\rightarrow l_m\rightarrow U$ 的价格是景点 $l_1,l_2,\cdots,l_m$ 的价格的总和：

$$\sum_{i=1}^{m}C_{l_i} \tag{3}$$

例如，行程 $U\rightarrow l_1\rightarrow l_5\rightarrow U$ 的价格为 10+100=110。

定义 12(行程的得分)　行程 $U\rightarrow l_1\rightarrow l_2\rightarrow\cdots\rightarrow l_m\rightarrow U$ 的得分是景点 $l_1,l_2,\cdots,l_m$ 的得分分别乘以群体相应偏好的总和：

$$\sum_{i=1}^{m}S_{l_i}\times p_{c_{l_i}} \tag{4}$$

其中，$p_{c_{l_i}}$ 表示群体对 l_i 所属分类的偏好程度。

例如，行程 $U\rightarrow l_1\rightarrow l_5\rightarrow U$ 的得分为 4.5 * 0.5+4 * 0.5=4.25。

定义 13(有效行程)　若行程 $U\rightarrow l_1\rightarrow l_2\rightarrow\cdots\rightarrow l_m\rightarrow U$ 的时间和价格分别满足群体的时间和价格约束则该行程是一个有效行程。

行程 $U\rightarrow l_1\rightarrow l_5\rightarrow U$ 的时间 7 和价格 110 满足其时间 8 和价格 200 的约束，因此该行程是一个有效行程。

定义 14(剪枝行程)　行程 $U\rightarrow l_1\rightarrow l_2\rightarrow\cdots\rightarrow l_m\rightarrow U$ 因时间超过群体时间约束或者价格超过群体价格而被剪枝，又或者包含一个剪枝行程，则该行程是一个剪枝行程。

例如，行程 $U\rightarrow l_2\rightarrow l_3\rightarrow U$ 的时间为 8.5>8，因此该行程是一个剪枝行程。

3　群体发现

目前，有很多聚类方法可以用于挖掘用户群体。K-means 是一个经典的聚类算法，凭借时间复杂度低和易于实现的特点，在机器学习和数据挖掘领域得到了广泛的应用。但是，当样本空间是非凸时，该算法容易陷入局部最优。谱聚类算法是一个基于图论的算法，可以作用于任意样本空间，并收敛于全局最优。在本文中使用上述两种聚类算法(K-means 和谱聚类算法)挖掘用户群体，然后根据群体的签到信息获取群体的偏好。

3.1　群体聚类

K-means 算法[18]是基于距离的聚类方法，本文中采用的距离是式(1)描述的两个用户的归一化签到向量的余弦相似度。令簇中心的个数 k 的取值在 2 到 $|U|/2$ 之间变化。对于每一个 k 值，K-means 算法一般有以下 5 个步骤。

(1)初始化 k 个簇中心

本文使用最大最小距离法选择 K-means 聚类的初始 k 个簇中心。首先随机选取一个用户的归一化签到向量作为第一个簇中心，然后依次选择与簇中心中所有向量的余弦相似度的和最小的归一化签到向量作为下一个簇中心，直到簇中心中向量的个数达到 k 为止。

(2)划分用户到不同的簇中心

得到 k 个初始簇中心之后，计算所有用户的归一化签到向量到 k 个簇中心的余弦相似度，并将这些用户划分到余弦相似度最高的簇中。

(3)取每一个簇中所有用户的归一化签到向量的平均值作为新的簇中心，根据新的簇中心重新迭代并将用户划分到不同的 k 个新簇中。

(4)依次迭代步骤(2)、步骤(3)，直到划分的群体的结果不再变化。

(5)计算给定 k 值的轮廓系数。

轮廓系数[20]是由 Peter 于 1987 年提出的一种评价聚类效果的方法。轮廓系数结合了凝聚度和分离度来评价聚类效果，在−1 到 1 之间取值，值越大表示聚类效果越好。轮廓系数的计算方法为：对于一个特定的 k 值，使用 K-means 方法对用户进行聚类；然后使用式(1)计算第 i 个用户与同一个簇中其他用户的相似度的平均值，记为 A_i，表示簇内的凝聚度；最后计算第 i 个用户与其余每一个簇内所有用户的平均相似度，并取平均相似度最小的簇对应的平均相似度作为簇间分离度，记为 D_i。对于第 i 个用户，其轮廓系数 Q_i 由式(5)得到。

$$Q_i=\frac{A_i-D_i}{\max(A_i,D_i)} \tag{5}$$

计算所有用户的轮廓系数的平均值作为当前给定 k 值聚类的整体轮廓系数，最后选取轮廓系数最大的值对应的 k 值作为簇中心的个数。

接下来介绍怎样使用谱聚类挖掘群体。谱聚类[19]是一种基于图论的聚类方法，一般包含以下 4 个步骤。

(1)构建表示用户关系的矩阵 M。将用户作为节点，用户间共同签到的关系作为边，构造无向图 G，G 中边的权重为两个用户的归一化签到向量的余弦相似度；然后，构造图 G 的邻接矩阵 M，M 的对角线元素值为 0。

(2)归一化拉普拉斯矩阵 N，$N=O-M$。其中，O 为度矩阵，是一个对角阵，对角线元素为相应的行(或列)的和。

(3)通过计算 N 的前 k 个特征值与特征向量，构建特征向量空间。

(4)利用 K-means 聚类算法对特征向量空间中的特征向

量进行聚类。此处的 k 值仍然使用轮廓系数的思想来确定。

3.2 群体偏好获取

每一个位置都有一个分类，根据群体中用户的共同签到位置的分类来计算群体的偏好向量 $P_U=\langle P_1,P_2,\cdots,P_{|c|}\rangle$，其中向量 P_U 中的元素 P_i 表示群体 U 对 c_i 类位置的偏好程度，使用式(6)来计算。

$$P_i=\frac{|\{\sum_{l\in U.L}\sum_{u\in U}\hat{CI}_{u,l}:l.c=c_i\}|}{\sum_{l\in U.L}\sum_{u\in U}\hat{CI}_{u,l}} \tag{6}$$

其中，$U.L$ 表示群体 U 中所有用户共同签到过的位置的集合，$l.c=c_i$ 表示位置 l 所属的分类为 c_i。

4 行程推荐

本文行程推荐算法的思想来源于 Lu 等人[14]的工作，在 Lu 等人工作的基础上加入了群体的行为偏好，得到了适用于向群体用户推荐旅游行程的推荐算法(见算法 1)。其中，第 1—3 行定义了 3 个变量，*validTripList* 用来存储所有的有效行程；*prunedTripList* 用来存储所有已被剪枝的行程；*tripList* 用来存储包含 i 个景点的所有行程，当 i 变化时，*tripList* 被清空(第 5 行)。第 4 行 i 从 1 循环到 n，表示包含任意个($1-n$)景点的行程都要被判断是否有效，这是一个 NP 完全问题。第 6—11 行得到包含 i 个景点的有效行程。第 7 行中 *trip.time*(L)用于计算群体当前位置在 L，走过行程 *trip* 所需要的时间；*trip.cost*()用于计算行程 *trip* 的价格开销；*trip.isContainPrunedTrip*()用于判断行程 *trip* 中是否包含已经被剪枝的行程。当满足上述 3 种条件中任意一条时，行程 *trip* 将被剪枝，将其加入到 *prunedTripList* 中(第 8 行)；否则，行程 *trip* 为一个包含 i 个景点的有效行程，将其加入到 *tripList* 中(第 10 行)。第 12—14 行用来判断包含 i 个景点的所有行程是否全部被剪枝，如果包含 i 个景点的所有行程中包含有效行程，则将其加入 *validTripList*(第 13 行)，然后计算包含 $i+1$ 个景点的行程；否则，算法终止(第 15 行)。由定义 13 可知，如果包含 i 个景点的所有行程都被剪枝，那么包含 $i+1$ 个景点的所有行程同样都会被剪枝。第 17 行根据群体的偏好向量 P_U 计算所有有效行程的得分，并根据得分降序排列行程，然后推荐 top N 个行程(第 18 行)。

算法 1 行程推荐算法

输入：群体 U 的偏好向量 P_U，群体的时间约束 TIME，群体的价格约束 COST，群体当前位置 L

输出：top N 个推荐行程

```
1. validTripList;/* 所有有效行程 */
2. prunedTripList;/* 所有已被剪枝的行程 */
3. tripList=∅;/* 包含 i 个景点的所有有效行程 */
4. FOR i=1 to n /* n 表示景点个数 */
5.    tripList=∅;/* 清空 tripList */
      /* Trip(i)表示包含 i 个景点的所有行程 */
6.    FOREACH trip in Trip(i)
      /* 如果行程超过时间或价格约束，或包含已被剪枝的行程，该行程被剪枝 */
7.       IF(trip.time(L)>TIME || trip.cost()>COST || trip.isContainPrunedTrip())
8.          prunedTripList.add(trip);
9.       ELSE /* 包含 i 个景点的有效行程 */
10.         tripList.add(trip);
11.   END FOREACH
12.   IF(tripList.size()>0)
      /* 有效行程加入 validTripList 中 */
13.      validTripList.add(tripList);
14.   ELSE
      /* 含 i 个景点的行程全剪枝，算法终止 */
15.      Break;
16. END FOR
    /* 根据群体偏好降序排列行程 */
17. sort(validTripList,PU);
    /* 返回得分较高的 top N 个有效行程 */
18. RETURN validTripList(N);
```

接下来，通过一个简单的示例来介绍行程推荐算法的执行过程。假设群体 U 为$\{u_1,u_2\}$，其共同签到景点为$\{l_1,l_3\}$，l_1 和 l_3 的分类分别是 c_1 和 c_3。通过式(6)计算得到群体 U 的偏好向量为 $P_U=\{0.5,0,0.5,0\}$。那么根据图 1 所示的景点地图，使用图 3 中的行程推荐算法向群体 U 推荐旅游行程。首先，计算包含 1 个景点的行程的时间、价格以及行程得分，如表 2 所列。

表 2 包含 1 个景点的行程

景点	行程	时间	价格	行程得分
l_1	$U\rightarrow l_1\rightarrow U$	4	10	$4.5*0.5=2.25$
l_2	$U\rightarrow l_2\rightarrow U$	5	90	$2*0=0$
l_3	$U\rightarrow l_3\rightarrow U$	4	70	$3.5*0.5=1.75$
l_4	$U\rightarrow l_4\rightarrow U$	3	120	$4*0=0$
l_5	$U\rightarrow l_5\rightarrow U$	5	110	$4*0.5=2$

由表 2 可知，包含一个景点的行程都满足 U 的价格和时间约束。

然后计算包含 2 个景点的行程的时间、价格以及行程得分，如表 3 所列。

表 3 包含 2 个景点的行程

景点	行程	时间	价格	行程得分
l_1l_2	$U\rightarrow l_1\rightarrow l_2\rightarrow U$	8	100	2.25
	$U\rightarrow l_2\rightarrow l_1\rightarrow U$	8	100	2.25
l_1l_3	$U\rightarrow l_1\rightarrow l_3\rightarrow U$	7.5	80	4
	$U\rightarrow l_3\rightarrow l_1\rightarrow U$	7.5	80	4
l_1l_4	$U\rightarrow l_1\rightarrow l_4\rightarrow U$	6	130	2.25
	$U\rightarrow l_4\rightarrow l_1\rightarrow U$	6	130	2.25
l_1l_5	$U\rightarrow l_1\rightarrow l_5\rightarrow U$	6	110	4.25
	$U\rightarrow l_5\rightarrow l_1\rightarrow U$	6	110	4.25
~~l_2l_3~~	~~$U\rightarrow l_2\rightarrow l_3\rightarrow U$~~	8.5	—	—
	~~$U\rightarrow l_3\rightarrow l_2\rightarrow U$~~	8.5	—	—
~~l_2l_4~~	~~$U\rightarrow l_2\rightarrow l_4\rightarrow U$~~	—	210	—
	~~$U\rightarrow l_4\rightarrow l_2\rightarrow U$~~	—	210	—
l_2l_5	$U\rightarrow l_2\rightarrow l_5\rightarrow U$	6.5	190	2
	$U\rightarrow l_5\rightarrow l_2\rightarrow U$	6.5	190	2
l_3l_4	$U\rightarrow l_3\rightarrow l_4\rightarrow U$	7	190	1.75
	$U\rightarrow l_4\rightarrow l_3\rightarrow U$	7	190	1.75
l_3l_5	$U\rightarrow l_3\rightarrow l_5\rightarrow U$	6	170	2
	$U\rightarrow l_5\rightarrow l_3\rightarrow U$	6	170	2
~~l_4l_5~~	~~$U\rightarrow l_4\rightarrow l_5\rightarrow U$~~	—	220	—
	~~$U\rightarrow l_5\rightarrow l_4\rightarrow U$~~	—	220	—

在表 3 中，由于景点 l_2 和 l_3 组成的行程的总时间 8.5 超过了时间上限 8，因此景点 l_2 和 l_3 组成的行程 $U\rightarrow l_2\rightarrow l_3\rightarrow U$ 和 $U\rightarrow l_3\rightarrow l_2\rightarrow U$ 都被剪枝，不再扩展到包含 3 个景点的行程

中。此外，由景点 l_2 和 l_4 组成的行程以及 l_4 和 l_5 组成的行程的总价格超过了 U 的价格约束，因此，相应行程也被剪枝，如表 3 中删除线标注所示。最终，只需要将景点 l_1 和 l_2、景点 l_1 和 l_3、景点 l_1 和 l_4、景点 l_3 和 l_4 扩展到包含 3 个景点的行程中。计算包含 3 个景点的行程，如表 4 所列。

表 4　包含 3 个景点的行程

景点	行程	时间	价格	行程得分
~~$t_1t_2t_3$~~	—	—	—	—
~~$t_1t_2t_4$~~	—	—	—	—
$t_1t_2t_5$	~~$U\to t_1\to t_2\to t_5\to U$~~	9.5	—	—
	~~$U\to t_1\to t_5\to t_2\to U$~~	9.5	—	—
	~~$U\to t_2\to t_1\to t_5\to U$~~	10	—	—
	~~$U\to t_2\to t_5\to t_1\to U$~~	9.5	—	—
	~~$U\to t_5\to t_1\to t_2\to U$~~	10	—	—
	~~$U\to t_5\to t_2\to t_1\to U$~~	9.5	—	—
$t_1t_3t_4$	~~$U\to t_1\to t_3\to t_4\to U$~~	10.5	—	—
	~~$U\to t_1\to t_3\to t_4\to U$~~	10	—	—
	~~$U\to t_3\to t_1\to t_4\to U$~~	9.5	—	—
	~~$U\to t_3\to t_4\to t_1\to U$~~	10	—	—
	~~$U\to t_1\to t_4\to t_3\to U$~~	9.5	—	—
	~~$U\to t_4\to t_1\to t_3\to U$~~	10.5	—	—
$t_1t_3t_5$	~~$U\to t_1\to t_3\to t_5\to U$~~	9.5	—	—
	~~$U\to t_1\to t_5\to t_3\to U$~~	9	—	—
	~~$U\to t_3\to t_1\to t_5\to U$~~	9.5	—	—
	~~$U\to t_3\to t_5\to t_1\to U$~~	9	—	—
	~~$U\to t_5\to t_1\to t_3\to U$~~	9.5	—	—
	~~$U\to t_5\to t_3\to t_1\to U$~~	9.5	—	—
~~$t_1t_4t_5$~~	—	—	—	—
~~$t_3t_4t_5$~~	—	—	—	—

景点 l_1，l_2 和 l_3、景点 l_1，l_2 和 l_4、景点 l_1，l_4 和 l_5 以及景点 l_3，l_4 和 l_5 组成的行程中包含了已经被剪枝的行程，因此不需要再计算其时间和价格就可以直接将其剪枝。景点 l_1，l_2 和 l_5、景点 l_1，l_3 和 l_4 以及景点 l_1，l_3 和 l_5 组成的行程的价格都超过了价格约束，也都被剪枝。因此包含 3 个景点的行程全部被剪枝。根据行程推荐算法可知，当包含 3 个景点的行程全部被剪枝时，包含大于 3 个景点的行程也都会被剪枝，此时，算法结束。最后，将满足群体 U 的时间和价格约束的行程按照得分降序排列，如表 5 所列。在表 5 的基础上可以向群体推荐 top 5 和 top 10 的行程。

表 5　行程排序

行程	行程得分
$U\to l_1\to l_5\to U$	4.25
$U\to l_5\to l_1\to U$	4.25
$U\to l_1\to l_3\to U$	4
$U\to l_3\to l_1\to U$	4
$U\to l_1\to l_2\to U$	2.25
$U\to l_2\to l_1\to U$	2.25
$U\to l_1\to l_4\to U$	2.25
$U\to l_4\to l_1\to U$	2.25
$U\to l_1\to U$	2.25
$U\to l_2\to l_5\to U$	2
$U\to l_5\to l_2\to U$	2
$U\to l_3\to l_5\to U$	2
$U\to l_5\to l_3\to U$	2
$U\to l_5\to U$	2
$U\to l_3\to l_4\to U$	1.75
$U\to l_4\to l_3\to U$	1.75
$U\to l_3\to U$	1.75
$U\to l_4\to U$	0
$U\to l_2\to U$	0

5　实验分析

本节首先介绍实验中使用的数据集，然后给出了实验设计、实验结果的评价方法以及实验效果。所有实验均使用 Java语言实现，运行在 i3 处理器、4GB 内存、Win7 操作系统的电脑上。

5.1　数据集

对于向群体推荐行程的问题，尚未有公开数据集可以直接应用，因此从新浪微博爬取了从 2011 年 12 月到 2014 年 6 月部分用户的实际签到数据。由于本文主要研究向群体推荐旅游行程，因此，只保留数据集中用户在上海的景点的签到记录，删除了用户在非景点的签到记录以及在其他省份的签到记录，并删除了少于 5 个用户签到的景点和签到景点少于 5 个的用户。最终，数据集包含 42 位用户在 51 个位置的 1486 次签到记录，如表 6 所列，部分签到记录的分布如图 3 所示。由于新浪微博中不包含景点的部分信息（如景点的分类、用户对景点的打分、价格、游玩时间），因此从大众点评爬取了景点的相应信息，获得了景点的分类、得分（0－5 分制）、价格、游玩时间等信息，从而形成了完整的实验数据集。群体的偏好是群体对景点分类的偏好，数据集中景点的分类来源于大众点评，主要包含：动/植物园、古镇、寺庙、教堂、游乐场、购物中心、郊游等。

表 6　数据集信息

类型	数量
用户	42
位置	51
签到记录总数	1486
每个用户的平均签到次数	35

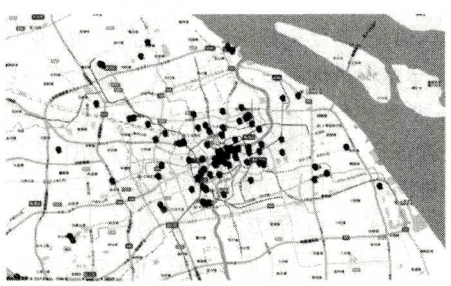

图 3　部分签到记录分布图

5.2　实验设计

在实验中，根据每一个用户签到数据的时间标签将用户的签到数据分为测试集和训练集。选取用户在某个时间标签前的 70%的签到数据作为训练集，该时间标签后的 30%作为测试集。根据用户在训练集上的签到记录，使用本文第 3 节介绍的方法聚类得到 7 个用户群体，并获取每个群体的偏好；然后模拟群体当前所在位置以及群体的时间和价格约束，使用第 4 节介绍的行程推荐算法向群体推荐旅游行程。本文使用 F 值来评价实验效果。

F 值＝（群体准确率 * 群体召回率 * 2）/（群体准确率＋群体召回率）

群体准确率＝群体中所有用户准确率的平均值

群体召回率＝群体中所有用户召回率的平均值

用户准确率=推荐命中的景点的个数/推荐行程中景点的个数

用户召回率=推荐命中的景点的个数/测试集中用户签到的景点个数

5.2.1 向群体推荐旅游行程

根据群体 F 值的定义,本实验分别向使用 K-means 聚类和谱聚类得到的群体推荐 top 5 和 top 10 个行程,得到的群体 F 值如图 4 所示。

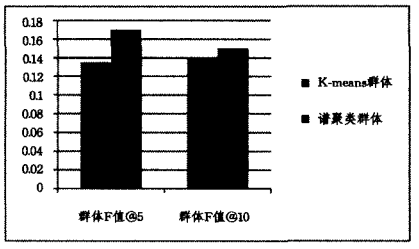

图 4 群体 F 值

由图 4 可知,通过本文提出的群体行程推荐方法得到的群体 F 值较高,并且使用谱聚类得到的群体的推荐效果比使用 K-means 聚类得到的效果更好。其原因是 K-means 算法建立在凸球形的样本空间上,当样本空间不为凸时,算法会陷入局部最优,且初始簇中心 k 的取值对 K-means 聚类算法的效果有较大影响。例如,用户 u_5 的归一化签到向量为(0.4,0.5,0,0,0);用户 u_6 的归一化签到向量为(0.7,0.7,0,0.4,0.8);用户 u_7 的归一化签到向量为(0,0,0,0.8,0.7)。此时,用户 u_5 和 u_6 的归一化签到向量的余弦相似度很高,并且用户 u_6 和 u_7 的归一化签到向量的余弦相似度也较高,那么用户 u_5,u_6 和 u_7 就可能被聚类到同一个群体中;但事实上,用户 u_5,u_6 和 u_7 并没有共同签到位置,此时计算群体的偏好是不准确的。但是,谱聚类算法能在任意形状的样本空间上聚类,且收敛于全局最优解,因此可以减少上述情况的发生。

5.2.2 基于用户向群体推荐行程

若设置每个群体只包含一个用户,问题就退化为向单个用户推荐旅游行程,分别向单个用户推荐 top 5 和 top 10 个行程,得到的用户 F 值为 22.6%和 21.5%。由此可见,向单个用户推荐行程的效果要优于向群体推荐行程的效果。但是,如果直接将向单个用户推荐的行程推荐给该用户所在的群体,推荐效果则较差。该实验首先对一个群体中的每一个用户进行行程推荐,然后将每一个用户的行程推荐给该群体得到一个群体 F 值,最后取所有用户推荐给该群体所得到的群体 F 值的平均值作为基于用户向群体推荐的结果,如图 5 所示。

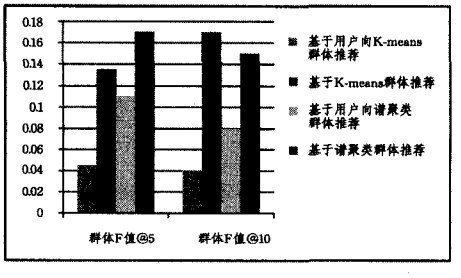

图 5 基于用户的群体 F 值

由图 5 可知,将向单个用户推荐的行程推荐给该用户所在的群体(K-means 和谱聚类)时,群体 F 值比使用群体偏好得到的群体 F 值更低。因为在向单个用户推荐的过程中只考虑了该用户对景点分类的偏好,而在一个群体中每个用户的偏好必然存在差异,所以不能依赖单个用户的偏好向群体进行推荐,这也进一步说明了向群体用户推荐旅游行程的问题具有一定的研究价值。本文提出的行程推荐算法是一个 NP 完全问题,随着位置个数的变化,算法的运行时间呈指数形式增长,如图 6 所示。

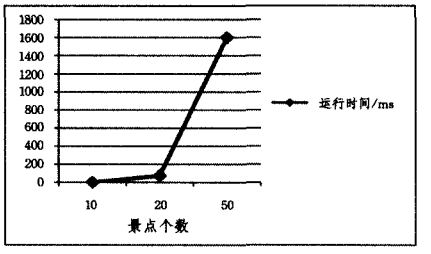

图 6 行程推荐算法的运行时间

结束语 本文通过用户的签到记录,使用 K-means 和谱聚类方法挖掘用户群体,并根据群体中用户的签到记录计算其偏好,然后综合考虑群体的时间和价格约束向群体推荐符合其偏好的行程。通过一系列对照实验验证了本文提出的技术路线的可行性和有效性。在未来的工作中,将考虑在群体聚类过程中加入用户间的社交关系,使得群体聚类更为准确;同时本文的行程推荐算法是 NP 难问题,如何寻找近似最优的行程推荐算法也是下一步努力的方向。

参 考 文 献

[1] ZHENG Y,ZHANG L,XIE X,et al. Mining interesting locations and travel sequences from gps trajectories [C]// International Conference on World Wide Web WWW 2009. Madrid, Spain,DBLP,2009:791-800.

[2] QUAN Y,GAO C,MA Z Y,et al. Time-aware Point-of-interest Recommendation [C]// SIGIR. 2013:363-372.

[3] YE M,YIN P,LEE W C,et al. Exploiting geographical influence for collaborative point-of-interest recommendation [C]// SIGIR. 2011:325-334

[4] ZHENG V W,CAO B,et al. Collaborative Filtering Meets Mobile Recommendation: A User-centered Approach [C]// Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI-10). Atlanta,Georgia,USA,2010.

[5] YOON H,ZHENG Y,XIE X,et al. Smart itinerary recommendation based on user-generated gps trajectories [C]// International Conference on Ubiquitous Intelligence and Computing. Springer-Verlag,2010:19-34.

[6] YUAN J,ZHENG Y,XIE X. 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. 2012:186-194.

[7] GARCIA I,SEBASTIA L,ONAINDIA E,et al. A Group Recommender System for Tourist Activities [C]// International Conference on E-commerce and Web Technologies. DBLP,2009: 26-37.

[8] ZHANG J W,KAWASAKI H,KAWAI Y. A Tourist Route

Search System Based on Web Information and the Visibility of Scenic Sights [C]// Second International Symposium on Universal Communication (ISUC 2008). 2008:154-161.

[9] CRANDALL D, BACKSTROM L, HUTTENLOCHER D, et al. Mapping the World's Photos [C]// Proc. Int. Conf. on World Wide Web (WWW). 2009:761-770.

[10] ASNBROOK D, STARNER D. Using GPS to Learn Significant Locations and Predict Movement across Multiple Users [J]. Personal and Ubiquitous Computing, 2003, 7(5):275-286.

[11] IWATA T, WATANABE S, YAMADA T. Topic Tracking Model for Analyzing Consumer Purchase Behavior [C]// Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI). 2009:1427-1432.

[12] HOFMANN T. Probabilistic Latent Semantic Analysis [C]// Proc. Conf. on Uncertainty in Artificial Intelligence (UAI). 1999:289-296.

[13] KURASHIMA T, IWATA T, IRIE G. Travel route recommendation using geotags in photo sharing sites [C]// ACM International Conference on Information and Knowledge Management (CIKM). 2010:579-588.

[14] LU E H C, CHEN C Y, TSENG V S. Personalized trip recommendation with multiple constraints by mining user check-in behaviors [C]// Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL). 2012:209-218.

[15] WANG H N, LI G L, FENG J H. Group-Based Personalized Location Recommendation on Scenic Sights[C]// Proc. Int. Conf. on APWeb. 2014:68-80

[16] HE W, LI D Y, ZHANG T L, et al. Mining regular routes from gps data for ridesharing recommendations [C]// Proceedings of the ACM SIGKDD International Workshop on Urban Computing(UrbComp'12). New York, NY, USA, 2012:79-86.

[17] LIN Y R, SUN J M, CASTRO P, et al. Metafac: community discovery via relational hypergraph factorization [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009:527-536.

[18] MACQUEE Q J. Some Methods for Classification and Analysis of Multivariate Observation [C]// Proceeding 5th Berkley Symposium on Mathematical Statistics and Probability. 1967: 281-297.

[19] SHI J, MALIK J. Normalized Cuts and Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8):888-905.

[20] ROUSSEEUW P J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis [J]. Journal of Computational and Applied Mathematics, 1987, 20(20): 53-65.

(上接第 198 页)

问题。引入处理大数据的思想(如 Map Reduce 算法等)来处理高维特征，是处理具有高维特征区间数样本的有效途径之一。

参 考 文 献

[1] MOORE R E. Interval arithmetic and automatic error analysis in digital computing[D]. Palo Alto: Stanford University, 1962.

[2] FILIPPONE M, MASULLI F, ROVETTA S. Applying the possibilistic c-means algorithm in kernel induced spaces[J]. IEEE Transcations on Fuzzy Systems, 2010, 18(3):572-584.

[3] REN S J, LV J H. Genetic algorithm based kernel function FCM clustering algorithm for interval numbers[J]. Journal of System Engineering, 2008, 23(5):611-616. (in Chinese)
任世锦，吕俊怀. 基于遗传算法的区间数核模糊聚类算法[J]. 系统工程学报，2008，23(5)：611-616.

[4] PIMENTEL B, COSTA A, SOUZA R. Kernel-based fuzzy clustering of interval data[C]// Proceedings of 2011 IEEE International Conference on Fuzzy Systems. Taipei, 2011:497-501.

[5] PIMENTEL B, COSTA A, SOUZA R. Input space versus feature space in kernel-based interval fuzzy C-Means clustering[C]// Proceedings of 2015 International Joint Conference on Neural Networks. 2015:1-7.

[6] VAPNIK V N. The Nature of Statistical Learning Theory[M]. London: Springer, 2000.

[7] TAX D M J, DUNI R P W. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20(11):1191-1199.

[8] SCHOELKOPF B, SMOLA A J. Learning with kernels: support vector machines, regularization, optimization, and beyond[M]. Cambridge, Massachusetts: The MIT Press, 2002.

[9] CAMPBELL C, BENNETT K P. A linear programming approach to novelty detection[C] // Proc of the Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, Canada, 2001:395-401.

[10] UTKIN L V, CHEKH A I. A new robust model of one-class classification by interval-valued training data using the triangular kernel[J]. Neural Networks, 2015, 69:99-110.

[11] CARVALHO F, SOUZA R, BEZERRA L. A dynamical clustering method for symbolic interval data based on a single adaptive Euclidean distance[C]// Proc of the Ninth Brazilian Symposium on Neural Networks (SBRN'06). 2006.

[12] KRISHNAPURAM R, KELLER J M. A possibilistic approach to clustering[J]. IEEE Transcations on Fuzzy Systems, 1993, 1(2):98-110.

[13] ANDERSON D T, BEZDEK J C, POPESCU M, et al. Comparing fuzzy, probabilistic, and possibilistic partitions[J]. IEEE Transcations on Fuzzy Systems, 2010, 18(5):906-918.

[14] CHEN B. Research on Outlier Detection Method and Its Key Techniques[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2013. (in Chinese)
陈斌. 异常检测方法及其关键技术研究[D]. 南京：南京航空航天大学，2013.

[15] HEIJDEN F, DUIN R, RIDDER D, et al. Classification, parameter estimation and state estimation-an engineering approach using Matlab[M]. Wiley, 2004.

[16] LICHMAN M. UCI Machine Learning Repository[DB/OL]. http://archive.ics.uci.edu/ml.