QQ 用户场景推断

KianXiao

June 23, 2016

Abstract

通过对用户上报数据的分析，我们首先提出了基于地点属性和用户访问时间规律推断用户场景的指导思想。接着我们对单个用户一个月的历史记录进行了分析，发现多数用户存在除家和公司外的其它场景，说明数据里面确实存在一些值得进一步挖掘的信息，但我们无法直接通过到访时间规律直接推断其场景。自然地，我们尝试提取地点属性信息，受数据的限制，我们提出了利用总体用户的到访记录对地点进行刻画的假设，实验中发现，通过这些记录所提取的特征不具备良好的区分性，通过多重分析，我们发现整体用户的上报规律淹没了地点本身的特性，假设不成立。受目前数据的限制，我们无法展开进一步有效的分析。

1 项目介绍

1.1 项目需求

目前我们拥有 QQ 用户每天上传的 GPS 位置信息以及对应的时间戳，希望挖掘出用户的一些常驻点信息，比如家，公司，娱乐场所等。

Table 1: 数据格式

id	session_lng	session_lat	place_id	session_start_time	session_end_time
1	113864371	22652942	11754545061	1449210820	1449210820
1	113864491	22652272	11754545060	1450321571	1450321571
1	113864614	22656695	11754545062	1449391447	1449391819
1	113864614	22656713	11754545062	1451464555	1451465121
1	113864614	22656713	11754545062	1451483039	1451483834
1	113864614	22656713	11754545062	1451485254	1451485254
1	113864614	22656714	11754545062	1451216370	1451217606

1.2 项目分析

我们先分析下用户的场景是由哪些因素所制约的，不难发现：

$$\text{地点属性} + \text{时间规律} \rightarrow \text{场景}$$

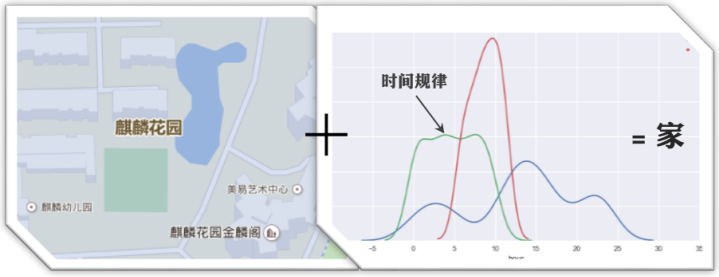


Figure 1: 示意图

也就是说用户在某个地点活动的时间规律和地点本身属性决定了用户所处的**场景**，这也是我们整个项目的指导思想。

很自然，我们可以将工作可以分为两个大的模块：地点属性提取，用户时间规律分析。

2 实验分析

2.1 用户时间规律分析

2.1.1 基于密度的聚类分析

根据用户一个月的历史位置信息，我们可以通过聚类算法挖掘出用户的常驻点信息，剔除掉一些噪声点。

我们先对数据做一个基本过滤，对**常驻点**作如下假设：对某一 Place_id 至少访问两次，并且最大访问时间间隔大于一天。

然后使用 DBSCAN 算法进行聚类，分析各类的访问时间规律。

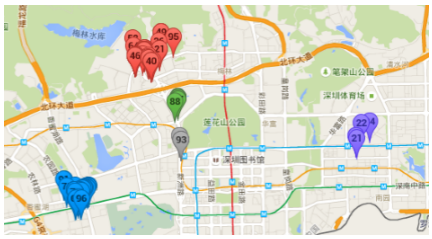


Figure 2: 用户聚类结果

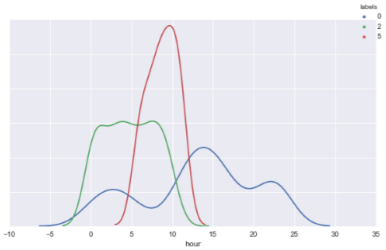


Figure 3: 各类访问时间规律

多数用户存在除家和公司外的其它场景，说明数据里面确实存在一些值得进一步挖掘的信息，但我们无法直接通过到访时间规律直接推断其场景。

我们还发现一些用户家和公司这两个场景无法有效区分（家和公司距离很近），如图2.1.1，集聚为一个团簇，我们希望在聚类分析的时候同时加上时间属性对 cluster 进行进一步划分。

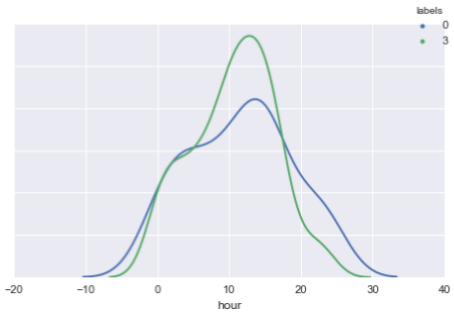


Figure 4: 家和公司难以区分的情况

2.1.2 改进的 DBSCAN 算法

为了解决上面的问题，我们对 DBSCAN 算法进行了改进 [1]，在邻域的判断里面使用了两个 metric，一个是空间距离，一个是时间距离。

我们可以得到一些有意思的结果，比如左边图里面的绿色曲线在凌晨段也有一部分凸起，和我们直观上的工作场景有一定的区别。我们加入时间属性再进行聚类的时候，这一类别就被划分为两部分，一部分和家的场景比较类似，一部分和工作的场景比较类似，说明该地点同时拥有用户的两种场景。

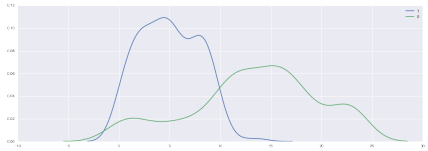


Figure 5: DBSCAN 聚类结果

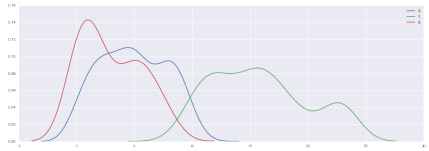


Figure 6: 改进的 DBSCAN 聚类结果

我们发现，公司和家这两个场景往往是大部分用户最强的两个场景，我们可以从时间规律上进行直接的推断，不需要结合地点属性信息。但是对于一些其他的场景，访问记录会明显减少，呈现出一定随机性，无法直接进行判断。

下面，我们需要对地点属性进行分析。

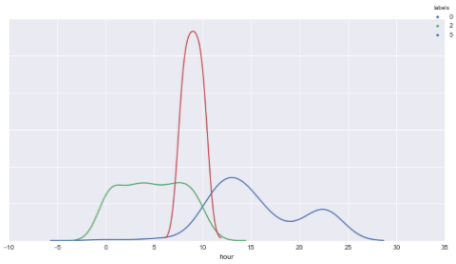


Figure 7: 其它场景

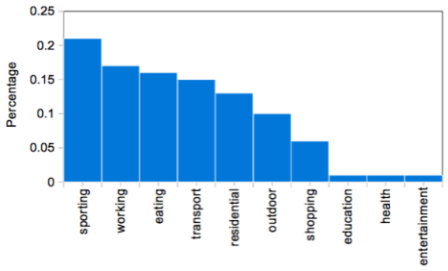
2.2 地点属性分析

2.2.1 根据 poi 分布

对于这个问题学术界普遍使用的是 FourSquare 上有标注的签到数据对地点进行聚类 [2]。

> 基本假设：一个地点的属性可以由该地点的 POI 数据所刻画。

首先统计每一个地点块上的 poi 分布，



将 poi 分布作为地点的特征向量进行聚类，然后得到每一类的 area profile，如图8、9。

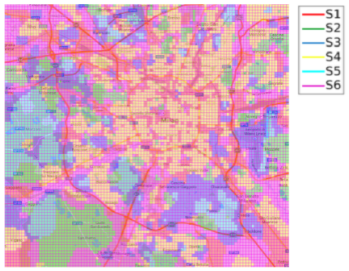


Figure 8: 聚类结果

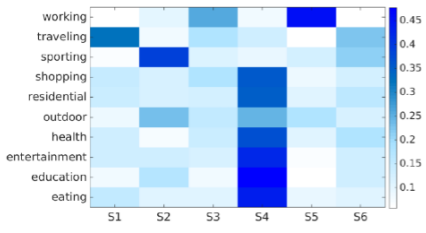


Figure 9: Area profile

这类方法思路很清晰，结果的可解释性比较强，但我们必须拥有相应的数据，这点不满足，其次，这种分析方法无法引入时间维度，地点的属性实际上应该是时变的，而不是固定不变。

2.2.2 根据用户行为信息

一个很自然的想法是，既然我们可以 poi 的分布去描绘地点的属性，那么我们能不能通过用户的行为数据对地点进行描绘呢？

基本假设：一个地点的属性可以由该地点用户的行为数据所刻画。

一个基本构想是，我们对每一个类别用一个 GMM 模型来描述，横轴为时间，纵轴为属于该类别的概率，如图2.2.2，对每一类地点进行解构 [3]。

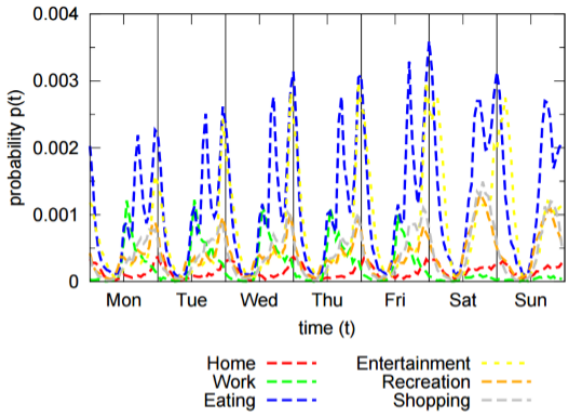


Figure 10: 各类别所对应的概率模型

根据每一个 place_id 内所有用户的行为特性，分解出每个类别的 GMM 模型，调研得到一种称为 I^2GMM 的方法 [4]。

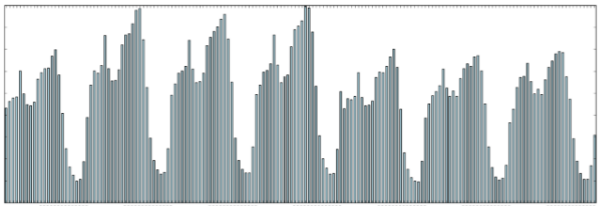


Figure 11: 某 place_id 上所有用户的上报规律

那么，要使用这套方法，首先我们得确保刚才的假设是成立的，也就是说，一个地点的属性可以由该地点用户的行为数据所刻画。

接下来我们根据每个 place_id 的用户行为数据所提取的特征进行聚类，观察各类别间的差异。

我们首先对提取的 48 维特征进行 PCA 降维，观察数据在三维空间的特点2.2.2。

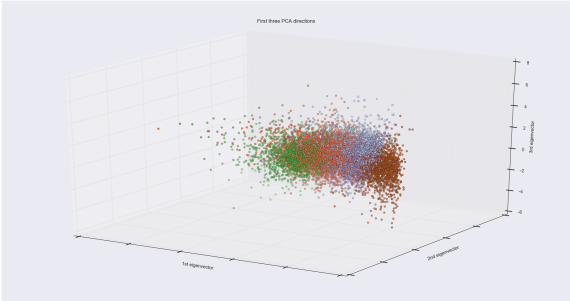


Figure 12: PCA 降维数据可视化

从 PCA 降维后的图我们可以发现，数据没有形成明显的团簇，不具备强区分性。我们使用该特征进行 Kmeans 聚类分析，做地图可视化进行进一步验证，从图2.2.2我们可以看出结果杂乱无章，难以从全局上进行合理化的解释。

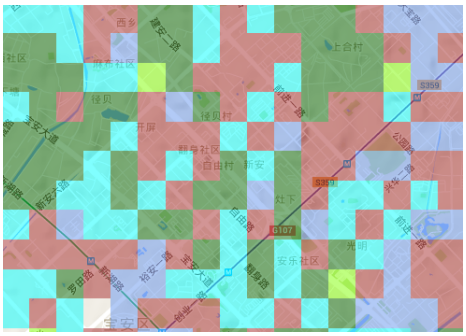


Figure 13: 聚类结果可视化

从这两份结果，我们可以从一定程度上推断，地点之间的差别被用户的上报规律所淹没。我们再来看下各类别对应的上报规律曲线，如图2.2.2

分析到这，我们基本可以确认，整体用户的上报规律（曲线走势）掩盖了地点的特性，我们难以通过上报规律对地点进行划分。

也就是说通过用户上报规律对 place 进行划分这条路无法进行。

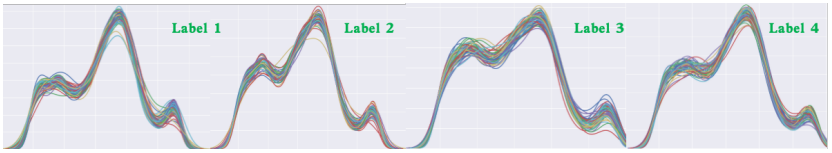


Figure 14: 各类别上报量随时间变化规律

2.2.3 可能情况分析

我们猜测可能是因为用户的两个主要场景：家和公司，掩盖了其它场景，所以我们对数据进行了过滤，过滤掉每个用户家和公司的数据，再做聚类分析，得出的结果依然不乐观，而且数据会变得很稀疏。

另外，我们对聚类后各类别 poi 分布进行了分析：

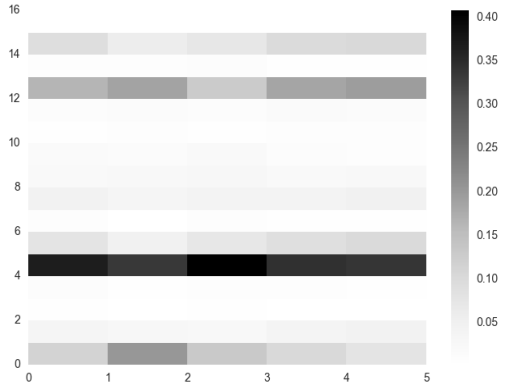


Figure 15: 各类地点 poi 类别分布

纵轴分别表示：公司企业，医疗保健，基础设施，娱乐休闲，房产小区，教育学校，文化场馆，旅游景点，机构团体，汽车，生活服务，美食，购物，运动健身，酒店宾馆，银行金融。

我们的 poi 数据量比较小，所以该分析图可能无法反应出真实的情况。

2.3 小结

我们首先确定了整体的分析思路，根据地点属性结合用户的到访时间规律对用户所在的场景进行分析。

在对单个用户的分析里面，我们通过聚类得到了用户的各场景所对应的模型，通过分析发现家和公司两种主要模式可以直接从用户的到访时间规律上进行区分，而其他的模式但从时间过滤上很难进行判断，于是很自然引出了地点属性。

在分析地点属性的过程中，我们首先调查了目前业界的常用方法，因为数据的限制，我们无法直接用 poi 信息对地点进行直接的聚类分析，所以我们想到利用用户的行为规律去表征地点之间的区别。接下来便是可行性分析，经过特征提取，聚类之后发现，各类别之间的差异性比较小，地点的特性被整体用户的上报规律所掩盖。

到此，我们可以得出结论，该方法无法很好的对地点进行划分，假设不成立。

3 总结

通过这个项目获取的最大收获应该是逐渐形成了自己分析问题的体系，由大到小，由简到繁。虽然最后没有得出一个好的结果，但这过程中尝试过的每一条路都收获颇多。

感谢 crazy 的耐心指导，帮我理清整个项目的思路，获取了不同的思维视角。

References

[1] ST-DBSCAN: An algorithm for clustering spatial–temporal data

[2] Characterization of behavioral patterns exploiting description of geographical areas

[3] Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media

[4] The Infinite Mixture of Infinite Gaussian Mixtures