一种商铺定位算法

小序

这是阿里天池算法大赛的一个很好的题目，可惜参与太晚了，设计了这种算法没赶上完全实现。如果有进入复赛的盆友感兴趣，可以试试，效果好的话，记着告诉我一声，让我也跟着乐呵一下。

——大桥之墩（670027200）

摘要

本文给出了一种商场内借助手机定位、WIFI 信息和历史交易数据定位商铺的算法。算法的关键在于将每个 WIFI 看作一个空间维度，计算当前连接的 WIFI 构成的多维空间点与候选店铺历史交易记录中最近的同维度空间点的距离作为评价指标，结合经纬度等商铺信息构成特征集，然后用随机森林模型进行预测。

问题描述

详见：

https://tianchi.aliyun.com/competition/introduction.htm?spm=5176.100066.0.0.2da3864eLsdt35&raceId=231620

问题分析

主要问题是如何有效利用 WIFI 的相互关系进行定位。

1. WIFI 是否连接对定位没有影响，所以不是特征；
2. 总共有近 100 个商场，每个商场内的 WIFI 环境互相独立，所以应该为每个商场建立一个独立的预测模型；
3. WIFI 中有位置比较固定的稳定 WIFI，也有用户自带的移动 WIFI，必须将不稳定的噪音 WIFI 过滤掉；
4. 根据稳定 WIFI 与商铺的关系，取用户能连接到的所有稳定 WIFI 关联商铺的交集，可确定候选商铺集合；
5. 观察候选商铺集合，发现有相同经纬度的商铺，说明不能根据 WIFI 确定楼层，以便用经纬度缩小候选商铺集合；可考虑统计商铺历史交易的最远经纬度差，过滤部分商铺；
6. WIFI 环境复杂，与商铺关联的稳定 WIFI 可能多达几百个，所以不能用模型直接学习商铺与关联稳定 WIFI 信号之间的关系；
7. 稳定的 WIFI 功率恒定，可以看着是一个空间的极坐标系统，这样空间的每个点就是多维空间点，当前 WIFI 强度构成的多维空间点和候选商铺历史同维空间最近点的距离可以构成评价指标；
8. 训练数据集应该有一部分数据参与构建历史多维空间库，一部分不能参与，测试数据则完

全不能参与；

9. 为了便于分析将 WIFI 信号强度变换为 0-100，历史多维空间缺维的，计算多维空间点距离时，该维度取最大距离即 100（注意不是整个距离取 100）；
10. 候选商铺数量不固定，所以需按候选商铺数量将 1 条记录拆分为多个样本，用二分类处理。

示例：

候选商铺中稳定 WIFI 的历史交易记录有：

{bssid=b001,signal=31;bssid=b003,signal=33;bssid=b005,signal=78}、

{bssid=b001,signal=32;bssid=b003,signal=45;bssid=b005,signal=60}、

{bssid=b001,signal=28;bssid=b003,signal=30;bssid=b005,signal=76;bssid=b009,signal=49}；

（1） 当前记录的稳定 WIFI：

{bssid=b001,signal=29;bssid=b003,signal=31;bssid=b005,signal=75}，则最小临距为：$\sqrt{(29-28)^2+(29-30)^2+(75-76)^2}$；

（2）{bssid=b001,signal=29;bssid=b003,signal=31;bssid=b005,signal=75}，则最小临距为：$\sqrt{(29-28)^2+(29-30)^2+(75-76)^2}$；

（2） 当前记录的稳定 WIFI：

{bssid=b001,signal=29;bssid=b003,signal=31;bssid=b008,signal=75}，则最小临距为：$\sqrt{(29-28)^2+(29-30)^2+(100)^2}$；

特征选取

特征包含 2 类信息：1、告诉模型候选商铺是哪家商铺的信息，包括候选商铺经纬度、商铺分类、均价；2、告诉模型行为发生时空间位置信息，包括用户的经纬度和 WIFI 信息。

1. 第一种选取方法

共选择 16 个特征，分别如下：

候选商铺经纬度、用户和商铺经纬度之差、商铺分类、均价、WIFI 信号 1—10 维最小距均值。少于 10 维的，填充 0。

2. 第二种选取方法：

共选择 9 个特征，分别如下：

候选商铺经纬度、用户和商铺经纬度之差、商铺分类、均价、WIFI 信号 N 维最小距、WIFI 信号 N—1 维最小距均值、稳定 WIFI 的数量。其中，N 为记录中稳定 WIFI 的数量。

3. 两种方案优缺点分析

第一种的特征数量较多，观察的维度也多，但是时间复杂度高；第二种只考虑最重要的两个维度，时间复杂度低，便于计算。

算法步骤及数据集划分

可按下列步骤实现算法：

1. 根据商场拆分训练数据；
2. 在每个商场的整个数据集上统计 WIFI 和商铺的关系，WIFI 参与商铺交易的次数；
3. 建立稳定 WIFI 集，如果 WIFI 参与某个商铺所有交易的 10%且次数大于 30，则为稳定 WIFI；
4. 在整个数据集中筛选基本数据集的记录：如果记录中所有的稳定 WIFI 关联商铺集取交集能唯一确定商铺的，则选出；

5. 将剩余数据等分为 3 份（这个比例可以调整）：训练 A 集、训练 B 集和测试集；
6. 用基本数据集和训练 A 集构建交易历史多维空间点库（为了方便而取的名字，实际上是以交易 ID 为外键的交易 WIFI 信号数据表，用 SQL 可取出对应记录）；
7. 建立训练和测试样本（注意要过滤掉噪音 WIFI）；
8. 训练和测试；
9. 统计结果（对于每条记录，选择候选商铺中为 1 概率最大的作为结果）。

小结

本文给出了一种商场内根据手机经纬度和 WIFI 间关联关系定位商铺的算法，提供一个思路，抛砖引玉，未经验证，欢迎有兴趣的朋友验证一下。