# 数据挖掘第一次作业

王倩妮 2015112956 交通 2015-02 班

## 问题一：影响地图匹配的因素有哪些?

有诸多因素对于地图匹配产生影响，我认为主要有以下几个方面。

首先，“GPS 的采样误差、偏差”会对地图匹配产生影响。GPS 采样质量又会存在以下几方面：

- ✓ 定位点位置的不精确、计算所得速度的不精确
- ✓ 数据传输过程中存在重复传输或者遗漏传输的状况，对于数据质量产生影响
- ✓ GPS 位点与采集时间上的不匹配(time stamp mismatched)
- ✓ 地图本身的精度状况，也会对地图匹配结果造成影响

其次，“数据采集的间隔时间”会对地图匹配产生影响。数据采集时的数据发送频率由 1s/次至 10 分钟/次不等，发送频率不同，对于匹配算法的选择存在影响，对于最终匹配结果的准确率也存在影响。准确率与时空开销是相互矛盾的两个变量，应根据数据用途、成本预算，选择合适的数据采集参数，并进一步选取合适的算法。

此外，“数据处理模式”会对地图匹配产生影响。按照数据处理的及时性，可将地图匹配问题分为“针对实时数据的地图匹配问题”与“针对历史数据的地图匹配问题”两类。选用不同的数据处理模式，t 时刻的下一刻 t+delta(t)数据存在“已知”与“未知”两种状态，数据处理的思路与依据不同。

接下来，针对数据处理模式，使用何种“地图匹配算法”对于地图匹配会产生影响。目前的地图匹配算法主要有以下几种分类：

- ✓ 基于几何关系的地图匹配算法（主要依据道路的几何关系，如距离、面积、角度等）
- ✓ 基于拓扑关系的地图匹配算法（主要依据路网的拓扑结构、连接关系等）
- ✓ 其他高级方法（如卡尔曼滤波、隐马尔科夫模型 HMM、机器学习……）

不同算法存在不同的适用范围与优缺点，使用何种地图匹配算法对 GPS 数据进行匹配，会对结果造成影响。此外，算法设计过程中的自身完整性、健壮性、稳定性也会对地图匹配造成影响。

## 问题二：如何改进算法?

对于算法的改进，需基于某一希望改进的算法，并在充分分析其不足的基础上进行。

对于课堂上的栅格化进行地图匹配存在“不易区分高架桥与平面道路”的问题与“不易区分平行路段”的问题。针对以上两种问题，提出两种尚未思考成熟的解决策略。

对于区分高架与平面道路，目前主要采用“限速区分”、“连续性区分”的手段。但有时限速无差异、拥堵状况的发生会减少以上两种区分方法的准确率。由于 GPS 数据主要依靠手机 APP 采集，且由于立交的遮挡会对于手机的信号强弱产生影响，我希望采用“信号强弱判别法”。首先判断某点可能匹配的路段中是否存在同时具有高架与平面路段的争议路段，若存在，则分析整段数据采集时的信号强弱，对应于争议路段的时间区域，判断此时手机信号有无明显阶梯状的衰减、或较长时间稳定于较低水平，从而进一步区分汽车行驶于高架道路还是平面道路。

对于平行道路的区分，若平行道路为不同等级道路，可以采用“限速区分”的方法。对于同等级的道路，暂时还未思考出合适的解决方案。

针对其他地图匹配算法，还有以下改进思路：

充分利用采集的各种数据信息，发挥几何、拓扑、道路属性信息等在地图匹配方面的优势，对几种匹配方式进行综合，确定不同方法应用过程中的权重参数，从而进行综合地图匹配，确定匹配结果。对于此种方法在应用于大型路网过程中效率低下的问题，可以使用机器学习过程中常使用的池技术，对于内存进行分配，以提高地图匹配的效率。

针对低频数据，采用隐马尔卡夫方法效果优于单纯几何方法，且对于最短路静的选择问题，可以选择更多启发式算法如 A-star 算法代替 Dijkstra 算法，进行算法的分步优化，加快最短路径的搜索速度。

## 问题三：如何正确评价算法?

对于算法要从不同的方面进行评价，使算法评价更为全面，为今后的算法改进打下坚实基础。

首先，在算法结果方面，要在“准确率”(accuracy)方面进行算法评价。本问题解决地图匹配问题，即将 GPS 记录点(GPS points)与路段(arcs/segments)进行匹配，结果即有某点“匹配成功”与“匹配不成功”两种状态，最终将结果连成一条路径(path)。对于地图匹配过程，可以选择某一组已知路径的测量数据（如公交车数据），统计其中匹配正确的 GPS 点百分比，作为准确率的衡量指标。因此，“准确率”是评价地图匹配效果、检验评价算法好坏的重要依据。

其次，在算法结果方面，要在“运行效率（时间）”(output delay/time)方面进行算法评价。不论是实时地图匹配算法(real-time algorithm)还是全局地图匹配算法(global algorithm)，不论是基于较长间隔、少量数据的算法，还是基于短间

隔、大数据的算法，都在一定程度上追求算法效率。算法效率与真正投入运营后的反馈效率与资金投入密切相关，因此研究者们也均在运行效率方面进行了考虑，在尽量短的时间内寻得尽量优的解。

此外，在算法结果方面，要在“运行占用空间”方面对于算法进行评价。充分考虑并压缩算法的时空开销，对于如今大数据背景下，数据的传输、存储、运算有着重要意义。也有研究在探索更好的数据存储方式，以最大限度降低调用难度，缩减运行占用空间。

接下来，在算法本身方面，要在算法的“适用范围”方面进行评价，不同算法由于推进思路不同，算法本身有一定适用范围与局限性，因此，衡量算法适用范围大小也是评价算法的一个重要因素。在地图匹配的过程中，不同算法存在不同的适用范围，例如有些地图匹配算法对于”U-turns”路段效果不佳，有些对于环形路段效果不佳，有些对于平行路段效果不佳，有些对于支路较多的路段效果不佳等等，因此评价算法也要考虑算法的局限性、适用范围大小。在应用过程中也应选择与问题应用区域更为吻合的算法，以保证更好的实施效果。

最后，在算法本身方面，要在算法的“健壮性”方面进行评价，GPS 数据存在一定误差，误差大小服从某一分布，算法对于异常数据的容错能力也是评价算法的重要指标之一。

因此，我认为可以通过以上指标的衡量，正确评价算法。