数据库专题训练大作业报告

2013011326 计32 周建宇

【问题描述】

大作业中给出了2009年5月1日、2日和3日三天总共1100万余个数据记录点，每一个点有一个它属于的路径id编号和其经纬度信息。我们需要从中找出所有相似的路径。两条路径相似的定义是A中任意一个采样点Ai与B中至少一点Bi的距离不大于100。

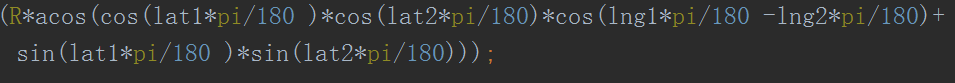
【核心算法】

该问题十分类似于第一次实验中的similarity search。不同的地方主要在于定义相似的方式不一样。Exp1中的定义方式为两个字符串的编辑距离不大于某一常数，或者jaccard相似度不小于某一常数。

在此我们采用了similarity search中的scancount算法作为寻找相似路径的核心方法。

首先我们先建立一张倒排表，作为第一层filter,具体为对于每个点，好到和它临近的点。之后，我们按照每个点的路径id进行分类汇总，当某一个点在某个路径id的集合中出现的次数等于该路径id所拥有的点数时，则证明该点所在的路径id与此当前路径id相似。

在这里相似度的计算为如下，其中R为地球半径，lati、lngi为第i个点的纬度和经度。



【数据结构】

与前三次实验类似，主要用unordered\_map存储索引，vector存储各种集合。

【注意要点】

1. 数据需要进行预处理，即将三个数据源文件合并为一个文件，并重写每个点的id编号。
2. 为了使处理速度尽量快，前期createIndex阶段要建立各种索引，不怕将相同的数据多存几份。用时为了防止数据对象拷贝过慢会直接用指针。