中国移动 China Mobile

5G+创新实训基地

技术实操的练兵场 · 能力认证的人才站 · 5G应用的孵化器

版权声明

中国移动 China Mobile

基于余弦和改进弹性距离的小区地址文本匹配

2020年11月

诺基亚-廖文哲

中国移动 China Mobile

Contents

项目简介与小区匹配-数据源处理

中国移动 China Mobile

项目简介：运营商家庭宽带地址都是现场人员手动输入，存在数据不准确，不全的现象；本项目通过文本匹配算法进行移动家庭宽带和杭州小区爬虫数据的匹配，统计各个小区宽带总数。

资管小区：default.D_RNT_IRM_IV_ADDRESSCOVER_D

覆盖场景	覆盖地域	中心位置经度	中心位置纬度	覆盖户数	关联资源点	标识	覆盖区域名称	地址	所属标准地址	所在小区/自然村/弄
coveragescene	coveragearea	centerpositionlongitude	centerpositionlatitude	usernumber	stronghold_id	id	name	addressinfo	address_id	address6_id
5	3	121.45257	29.98174	1	2793779	53941187	古巷社区	宁波市宁海县深圳镇环城东路8******	13883829	3589089

爬虫小区：

1.贝壳：appfx_ns2_hive_db.beike_list　　2.搜房网：appfx_ns2_hive_db.sofang_list

3.安居客：appfx_ns2_hive_db.anjuke_list　4.O域收集的全息小区。

city	point_name	county_name
杭州	港龙商业广场	江干
杭州	马塍路28号	西湖
杭州	逸城	富阳
杭州	凤起商务大厦	下城
杭州	浙江省通信产业服务有限公司	上城

预处理操作：

- 1.由于绝大部分爬虫小区point_name字段只有小区名，但是部分小区名会有诸如：杭州市XXX街道，(光纤箱)的字样，影响字符匹配，故对这些脏数据剔除。
- 2. 资管小区addressinfo字段很多小区名有（光纤箱XX号）脏数据，剔除这些脏数据以提高字符匹配准确度。
- 3.对爬虫小区根据经纬度和小区名去重。

小区匹配-算法介绍与改进

01

1.余弦相似度

在两份文本相似度上，“余弦相似度”算法被广泛使用。通过将两个文本出现的单词，建立A、B两个向量，并且计算这两个向量的余弦值。其数学表达为公式1：

$$Cos\_Similarity(A,B)=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|} \quad (1)$$

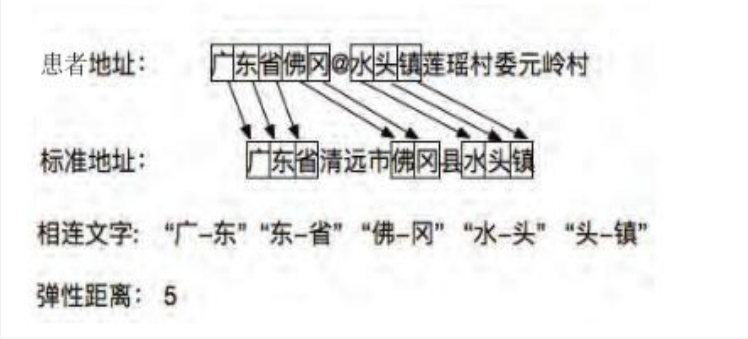
02

公式中A、B为文本编码后的向量；cos_Similarity(A,B)∈(0,1)代表了为两个文本的相似度，**当两份文本越相似，越接近1。**

2.弹性距离算法

由于余弦相似度算法本身是在向量空间里面计算两份文本的相似度，并不考虑字符的前后次序关系。为了表示这一特性，并且最大程度从患者填写地址中提取连贯的片段信息，定义“弹性距离”这一数学量，其概念表达见下图

将爬虫小区地址映射到资管地址相同的文字上，再检查标准地址上有多少个文字是相连的，最后数出相连的文字对数，标记为　弹性距离。若有多种映射方式，则选取最大的作为映射距离，其具有同时表示文字的次序和文字的紧密程度的优点。即便患者写出“北京市朝阳区广州市”的地址信息，也能正确地根据“北京市朝阳区”这段紧密相连的文字得到更大的权重，排除“广州市”相对于标准地址无关的信息干扰。弹性距离数学表达设患者地址文字序列为A=a 1 a 2 …a i，标准地址序列为B=b 1 b 2…bj，患者地址文字第l块片段Cl=cl1cl2…clk，其中C l ∈A且C l ∈B，则弹性距离为公式2：



小区匹配-算法介绍与改进

中国移动 China Mobile

03 3.算法改进：子集弹性距离

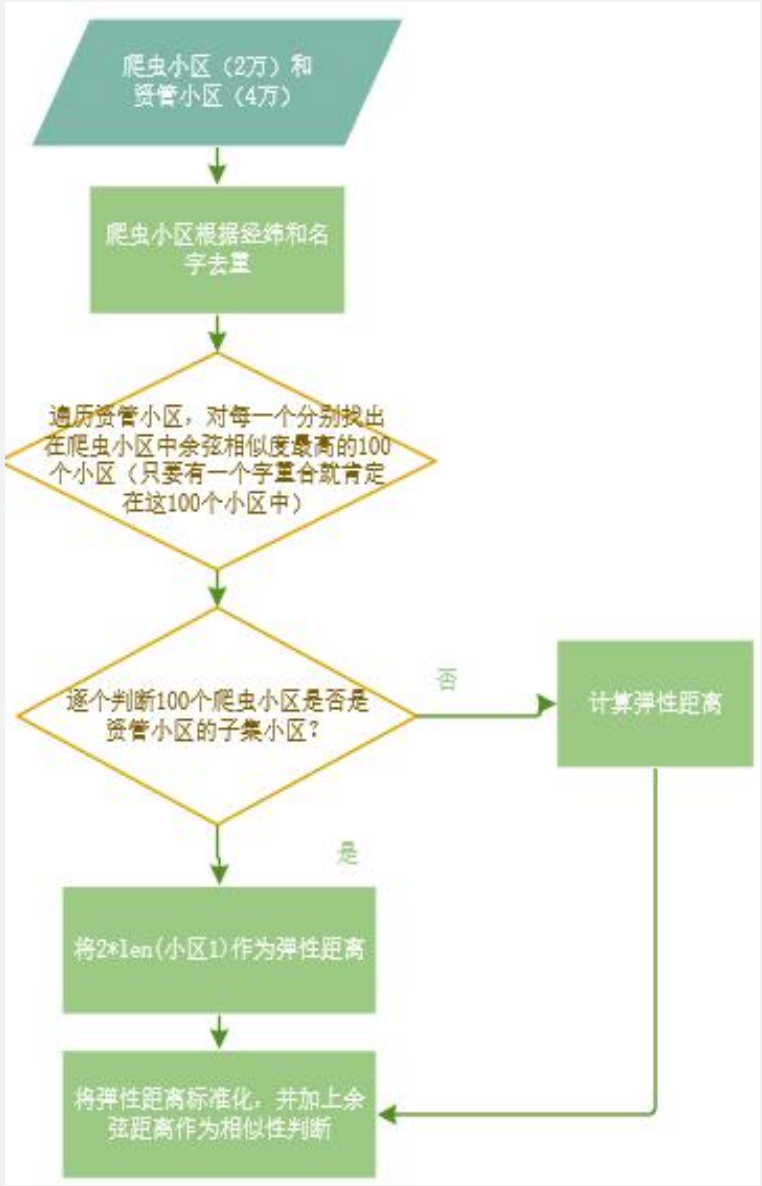
在资管和爬虫数据集上，传统的弹性距离的表现还有待改进，例如：

数据集1：玫瑰小区　花园北路20号玫瑰小区

数据集2：花园北路30号兴华小区　花园北路20号玫瑰小区

在数据集2上的弹性距离更大，代表两个小区更相似。但是实际上数据集1才应该作为最终结果。因此对传统的弹性距离改进，定义为：子集弹性距离，算法流程如下：如果小区1是小区2的子集，则定义两者之间的距离为2*len(小区1)，否则判断两者的弹性距离。

通过加大子集的权重，可以提高在资管和爬虫数据集上的算法准确度，经过测试，能达到5%左右。



04 4 结合各个算法的地址综合得分

由于余弦相似性算法不能考虑两个字符集之间的先后顺序，而子集弹性距离计算量过大，故将多个算法整合作为距离度量：计算爬虫地址对于每个资管地址的余弦相似性，将结果按高分排在前面，取前100个相似的地址，再对100个地址分别计算子集弹性距离，并将子集弹性距离的得分在100个地址的组内**最大最小归一化**：x ‘ = (x - X_min) / (X_max - X_min)（因为余弦距离的范围为[0,1]，Z-score规范化不能将取值范围严格控制在[0,1]），再与余弦相似性相加作为100个地址的最后得分，取最高分作为映射地址：

option address =Head10_address(Cos_Similarity(爬虫地址，资管地址表)

normallize score =normallize(D(option address))

best_ address =max(Cos_Similarity+normallize score)

小区匹配-整体匹配流程

中国移动 China Mobile

05

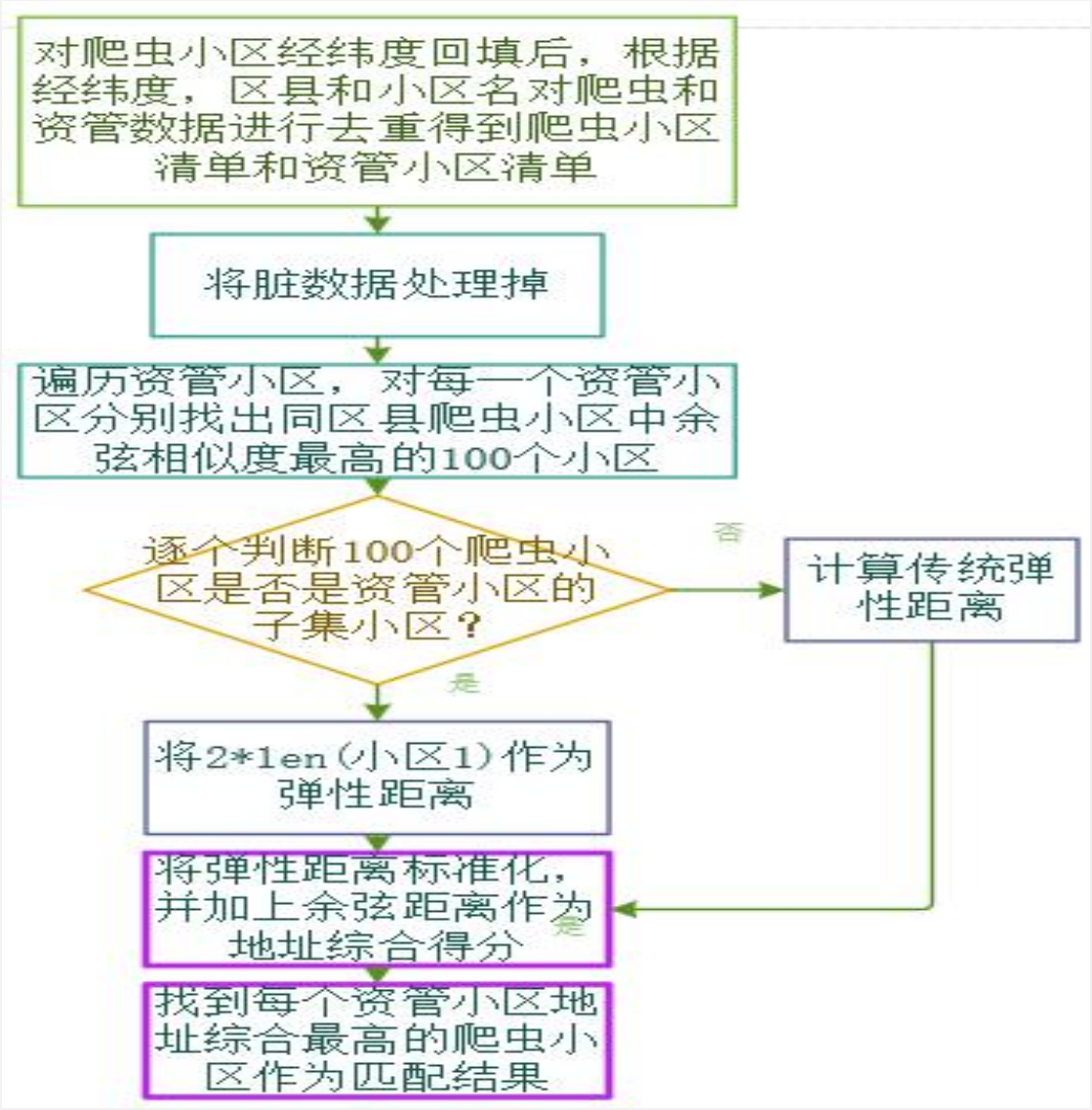
5.整体匹配流程

Step1.接入数据处理完后的资管小区和爬虫小区。

Step2.遍历资管小区，对每一个资管小区分别找出同区县爬虫小区中余弦相似度最高的100个小区，

Step3.逐个判断余弦相似度最高的100个爬虫小区是否是资管小区的子集小区,如果是，将2*len(爬虫小区)作为该爬虫小区的弹性距离，否则计算传统弹性距离，将传统弹性距离作为该爬虫小区的弹性距离。

Step4.将弹性距离标准化，并加上余弦距离作为地址综合得分,找到每个资管小区地址综合最高的爬虫小区作为匹配结果。



小区匹配结果统计与分析

结果统计

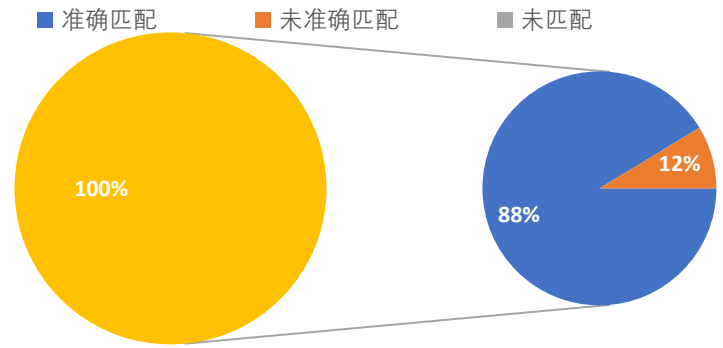
随机抽取资管杭州市三批每批200个样本数据先进行纯地址数据过滤然后进行人工的准确度测试

第一批：滨江资管小区总数：137　匹配准确数：125　匹配准确率：91%

第二批：上城+西湖资管小区总数：141　匹配准确数：117　匹配准确率：83%

第三批：滨江+西湖（苑，小区，社区筛选）资管小区总数：181　匹配准确数：164　匹配准确率：91%

总平均准确率：88%



结果分析

匹配准确率

针对未匹配成功的50条地址数据进行分析:

- **原因1**：发现46条（占比92%）地址数据爬虫库没有这个小区。
- **原因2**：发现4条（占比8%）算法误差导致。