《集合论与图论》大作业
疫情相关数据的获取、统计及分析

1190201215 冯开来

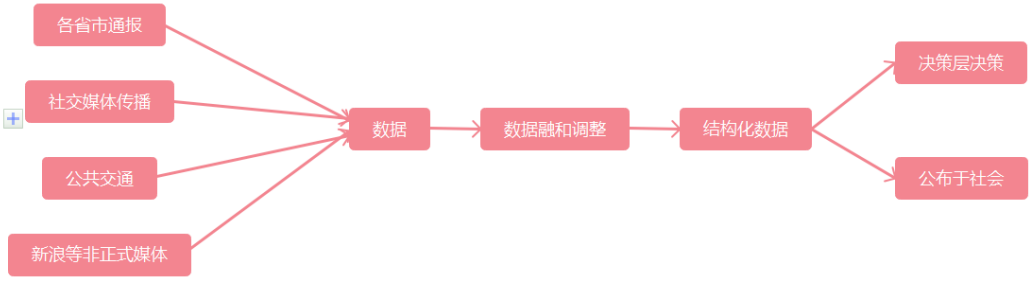
一、 研究背景

2020 开始之际，一场突如其来的疫情席卷了华夏大地。为了实时提供疫情情况，确诊疑似病例的数据统计显得格外重要。

有了准确的数据才能有效的为国家卫健委提供理论基础来进行全局的预测了解和分析调控。因此我们需要的数据越多越好，数据越精确越好。但是，疫情的扩散涉及到很多方面，比如各省市区通报的确证疑似病例，比如飞机火车汽车等公共交通会增加病毒的扩散面，比如病人隔离前去过的哪些场所，此外还有统计的时候要充分保证个人隐私……

二、 可能遇到的问题难点及分析

我们先建立一个流程，如图：



通过图解，我们列出以下一些难点和问题

1.如何确保要有广大的数据源

2.数据获得的可靠性，并且要过滤一些虚假信息和重复信息

3.因为疫情的扩散性，需要知道相关人员的最近 14 天的运动轨迹和密切接触者

4.数据要保证个人隐私不能泄露

5.相关接触者的运动轨迹等信息

6.每日核算检测量满足不了大量人员的检测

……

三、 应用价值

在疫情期间，数据统计分析的模型可以让每一个了解疫情动态，权威的如霍普金斯大学对全美甚至全球做出的数据统计和分析。

在疫情过后，这个模型还能利用在以下方面：

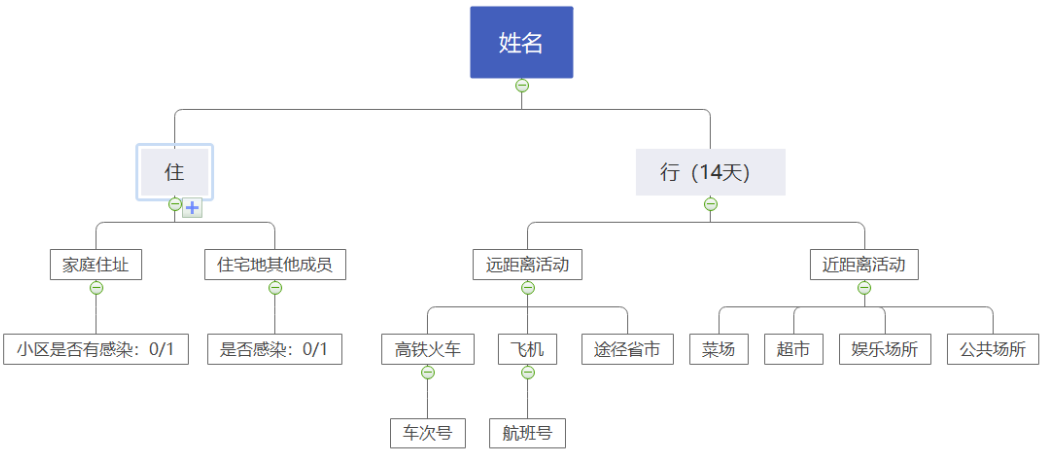
1. 猎头公司的人才资源管理
2. 人口普查
3. 军籍学籍的登记管理
4. 外国来华人口管理

……

四、 几个模型建立

1） 个人信息的图模型

首先，每个人的信息可以建立一个集合。这里我们通过一个人的住行来建立信息系统。如图：

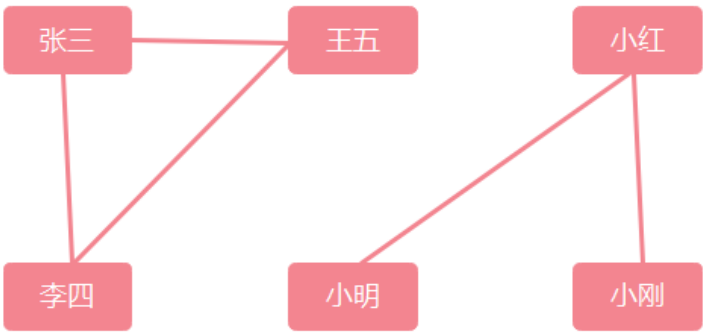


因为能力有限，暂列举出以上信息，现实生活中肯定更加复杂。我们这里通过几个简单的方面举几个个人信息例子。

{张三，0，A 小区，G123，H456，{M，N，O，P}}
{李四，1，B 小区，G123，H789，{O}}
{王五，0，B 小区，G123，H456，{M，N，P}}
{小明，0，C 小区，G011，H020，{R，S}}
{小红，0，C 小区，G011，H030，{S，T}}
{小刚，0，D 小区，G000，H030，{T}}
可以绘制出表格：

姓名	是否感染	住址	列车次号	航班号	活动场所一	活动场所二	活动场所三	活动场所四
张三	0	A	G123	H456	M	N	O	P
李四	1	B	G123	H789	O			
王五	0	B	G123	H456	M	N	P	
小明	0	C	G011	H020	R	S		
小红	0	C	G011	H030	S	T		
小刚	0	D	G000	H030	T			

因为新冠病毒的传播性极强，本着不能放过一个的原则，我们以个人姓名为顶点，住行方面如果有重叠则建立一个二元关系，这样我们可以得到一个具有若干无向边的无向图：



显然，这是一个具有若干划分的图，每一个支为一个划分。

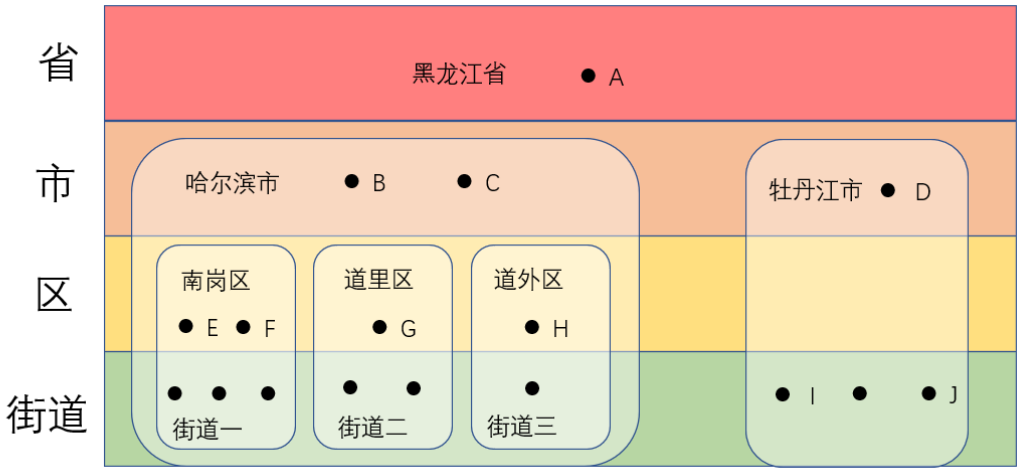
这种模型的优势在于能够快速将人进行分类，这样在核算检测的时候可以进行抽样检测，即形成一个相异代表系。从每个划分中抽部分进行检测，可以大大减少工作量。其次，每个人的信息都相关联，就如图中李四确诊后，我们可以快速锁定王五和张三进行检测或隔离。同时张三和王五同样关联着一些人，我们可以迭代重复，直至没有新的确诊病例出现。

2） 信息流的安全格模型

在各个省市区进行汇报的时候，这是一个自下由上的过程。这个可以简单的抽象为一个树图。这里我们将条件复杂化，假设每个平级省市区之间不能互相交流信息。同时，如果上级将信息传递个下级，那么下级将会知道他不应该知道的信息，会造成信息泄露，所以很显然，我们这里要建立一个具有偏序关系的链，但是因为信息有共同的信息源头和信息接收者，猜想这是一个格的模型。

因此这个模型可以适用于军用系统和政府系统，具有很严格的保密性和安全性。

这里我们以黑龙江省市区汇报作为例子，如图：



每个人的权力范围由集合表示：

H：{街道三}

B：{街道一，街道二，街道三}

D：{牡丹江市}

I：{牡丹江市}

A：{街道一，街道二，街道三，牡丹江市}

同时，同一部门的不同成员可能有不同权限

D：{牡丹江市}第三等级

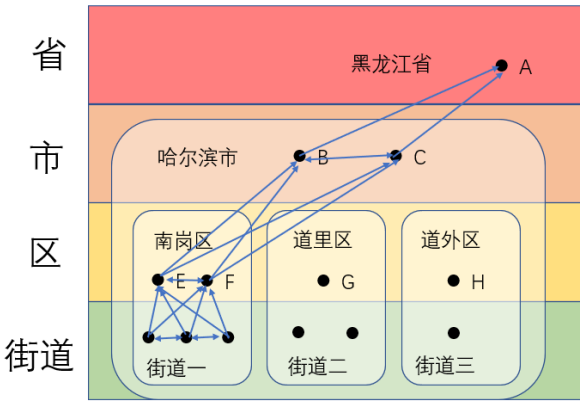
I：{牡丹江市}第一等级

不同部门的不同成员可能等级相同

H：{街道三}第三等级

E：{街道一}第三等级

G：{街道二}第三等级



因此，我们建立一个具有有向边的树模型（表示偏序关系）：

这里暂时只连线仅其中一部分，剩下以此类推。这里表示信息的流动只能沿着有向边的方向。

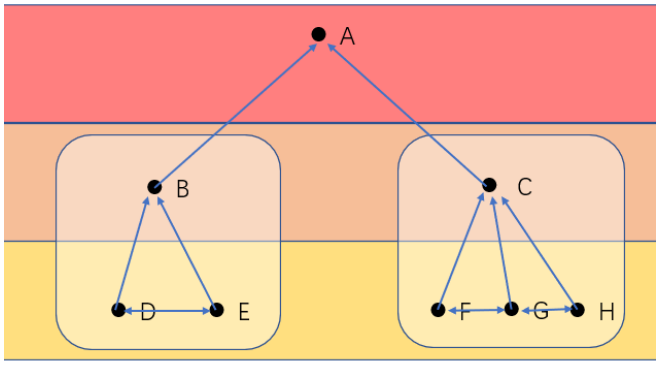
我们进行抽象，不同的权力范围可以看成彼此不相交的集合，用子集格 L_1 表示，不同级别我们用数值比较的线性格 L_2 表示。

对于权限的描述用线性格和子集格的积

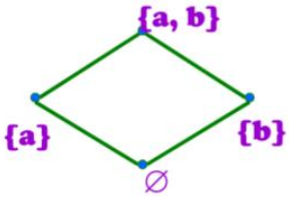
$$L_1 \times L_2$$

$(a,b) \leq (c,d) \cong a \leq c且b \in d$即在同一部门且级别低，这样信息才能流动。

我们继续简化模型，如图：

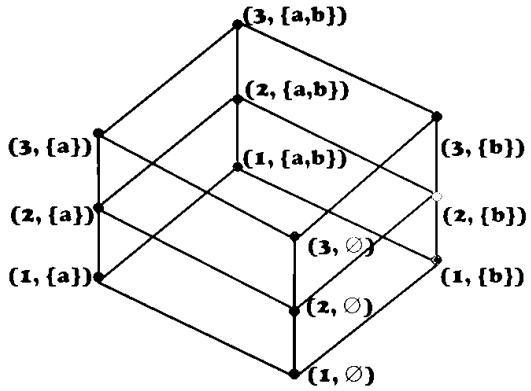


a



b

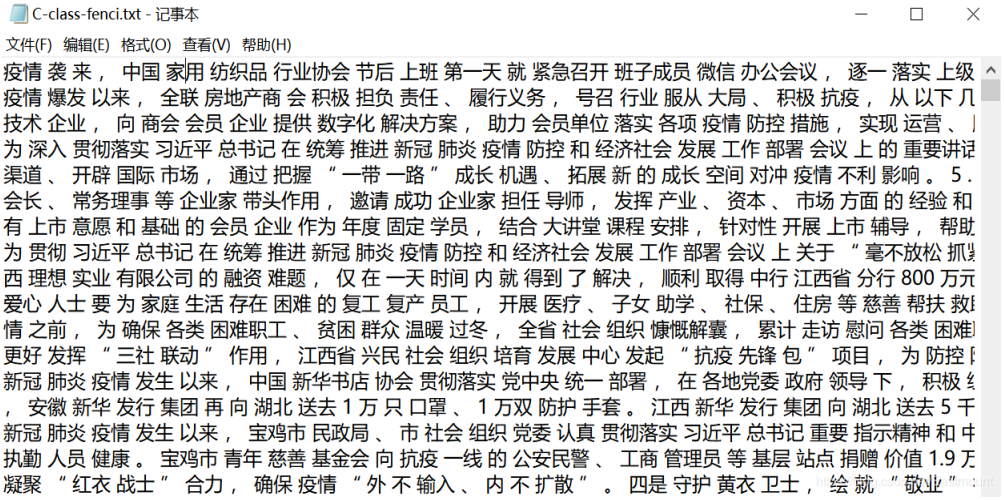
我们画出来权力范围 L_1 和不同级别 L_2 的格模型和链模型。因此二者的积格描绘了所有信息的流动方向和权限：



这样的格模型可以简化信息流动的情况，更加直观和简介。

3） 数据抓取和层次聚类的树模型

第一步肯定是分析网站，可以做一个爬虫进行数据挖掘，（因为本人还不会 python，所以这里只提供想法），在对每一个新闻标题进行中文分词，每一行代表一条新闻，并生成对应的内容，这里我们举个例子，如图：



然后我们对生成的词汇进行统计，得到一个词汇表，输出结果如左图所示，可以看到“疫情”、“组织”、“捐赠”、“社会”等都是高频词，也是我们老百姓关注的主题。现在我们对这些词汇进行层次类聚，我们不妨设出现次数为每个词汇到一固定点的距离。

名次	词汇	次数
1	疫情	20996
2	防控	15598
3	组织	14632
4	工作	13895
5	社会	12787
6	协会	11524
7	捐赠	10023

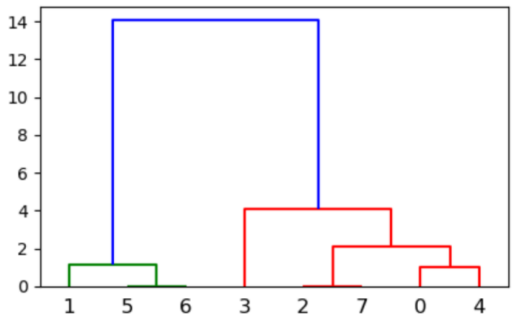
1 我们先计算 n 个对象两两之间的距离

2 在构造 n 个单成员聚类 $C1$，$C2$，……，Cn，每一类的高度都为 0。

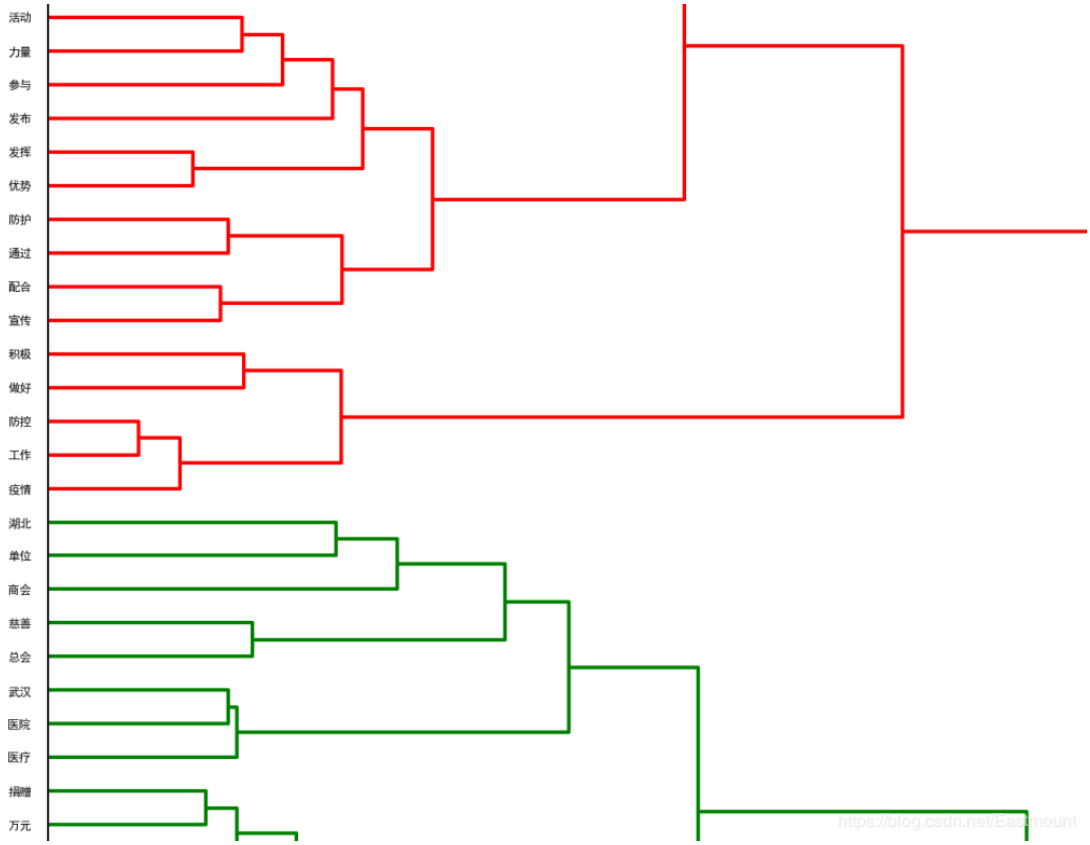
3 找到最近的两个聚类 C_i，C_j，合并后聚类减少 1，已合并的两个类间距作为上层的高度。

4 计算新生成的聚类与本层中其他聚类的间距，如果满足终止条件，算法结束，否则转 3。

如同这样：



结合疫情关键词，最终我们可以得到这样一个树（部分截图，原图太大了）：



层次类聚树的优点在于可以通过设置不同的相关参数值，得到不同粒度上的多层次聚类结构。并且可以随时停止划分。（这一块我了解的还不是很多，也没办法深入研究）

五、 结合个信息进行筛查（初步想法和体会）

基于以上几个模型，我们可以大致将整个过程氛围数据爬取，统计，分析，策划等过程。

1）首先利用信息流的安全格模型和爬虫的爬取获得确诊信息，获得确诊病例的数目统计和病人的信息系统

2）其次通过确诊病例的信息系统和个人信息的图模型进行筛查，通过不同的划分决定哪些人需要进行核酸检测，并且可以通过所形成的相异代表系选择性的进行检测，减少工作量和复杂度

3）最后通过层次聚类树决策出现在的舆论导向和热点问题，这样发布的文章可以获得更高的访问量

但是实际生活中的复杂程度远远高出我的理论模型，每个人的个人信息也不仅仅只有这么几项，并且每个人的行动轨迹遍布的范围更加广，涉及到的数据更加多，那个时候我们不仅需要找到求解模型还要简化运算量和复杂度。所以我在本文能做的就是简化模型的建立并且简单的求解模型。