

社会关系影响力模型（第一版）

权重因子构成因素(X1 feature待增加)。

权重因子用 w 表示，如果用户A与用户B是邻居节点，那么权重因子由下面因素构成。

- 1. 财富能力 q

$$w_{AB} \propto \frac{q_A}{q_A + q_B}$$

可能影响的人

- 1. 用户A对用户 a_1 的影响力：

$$\frac{1}{w_{Ai}} = \sum \frac{1}{w_{mn}}$$

w_{mn} 为从用户A到用户 a_i 最短路径上的权值。

Y值设计

- 1. 用户从A出发可到达的所有节点集合 $M = \{a_1, a_2, \dots a_n\}$
- 2. 用户A的总体影响力

$$Y = \sum_{a_i \in M} w_{Ai}$$

X2(feature待增加)

- 1. 职业
- 2. 收入
- 3. 城市

社交圈影响力

- 家庭财富分
 - 1. 家庭成员数量
 - 2. 家庭总财富
- 公司财富分

1. 公司员工数量
2. 公司员工总财富（财富分布可能有偏，待参考）
3. 企业年金总额

社会关系影响力模型（第二版）

新增问题及解决思路：

1. 通过社交网络补全联系方式

(1) 基于强关系的补全

- 家庭关系，同一家庭可以找到其他家庭成员的联系方式。
- 同事关系，同一公司可以找到同事的联系方式。

(2) 基于社交网络平台的补全

- 手机号和微博账号相互匹配，可以通过寻找较强社交关系的用户补全联系方式。
- 社交关系的衡量：
 - 微博上互相关注。
 - 相互评论较多。
 - 相互点赞较多。

2. 基于社交圈的推荐

(1) 基于明显的社交关系的推荐：

- 同一家庭选择影响力最大的人进行推荐
- 同一公司选择影响力最大的人进行推荐

(2) 基于相似用户的推荐（user-based）

- 历史购买记录相似。
- 财富，职业所在城市等等用户的上下文信息相似。
- 微博目前有用户的标签数据，比如#二次元#，#程序猿#，#股市大牛#，#网红#等等，可以用于寻找相似用户。
- 方法选择：基于历史购买记录的MF算法（可解释性较差）或者是socialFM算法。

3. 影响力权重feature的选取（待增加）。

需要更多对业务的理解

财富 职业 城市 公司实力

4. 可影响人群问题

(1) 单个用户的可影响人群：

用户的可影响人群列表就按照原来的方案计算，按照影响程度的大小进行排序。

(2) 社交网络影响力最大化

由于不可能对所有影响力大的人群进行重点营销（有可能这些影响力大的节点属于同一个群体，覆盖的人群重复率较高）。因此需要最大化整个社交网络的影响力。

基本思想就是：如何选取 k 个种子节点进行信息的传播和扩散，使得最终被激活的节点个数最多。

可供选择传播模型：

1. 独立级联模型

- 以发送者为中心，只能给自己的邻居节点传播一次。
- 初始化一系列的点作为开始的活跃节点，每一个活跃节点都有一次激活邻居节点的机会。激活成功的概率为 p_{ij} 。此时原来的节点还是活跃的，但是已经失去了激活其他节点的能力。邻居节点成为了新的活跃节点，重复之前的动作。知道图中没有拥有激活能力的活跃节点或者全部节点都变成的活跃节点时结束。
- 特点是它仅仅考虑 v 与出边邻居 w 之间的激活关系，完全不考虑 w 的其它入边邻居对 w 的影响。

2. 线性阈值模型

- 以接受者为中心，其他节点对自己的影响会累加。
- 每个节点都会被初始化一个阈值，表示该节点能够被影响的阈值。节点A对节点B的影响力必须超过B的阈值，B才能被激活。如果没有超过，A对B的影响力会被保存，如果存在节点C使得 w_{CB} 大于阈值 $\theta_B - w_{AB}$ ，那么B就会被激活。重复上述步骤，当网络中已存在的所有活跃节点中任意活跃节点的影响力之和都不能激活他们的处于非活跃状态的邻居节点时，传播过程结束。
- 当一个激活节点 w 尝试去激活它的未激活邻居节点 v 而没有成功时，节点 w 对节点 v 的影响力会被“积累”下来，而不是被抛弃，这种积累对后面节点 v 的其它邻居对 v 的激活时有贡献的，直到节点 v 被激活或传播过程结束。

注：虽然还有其他的传播模型比如pagerank,热力模型等等，但是没有以上两张基本模型符合我们的场景。

求解思路：

1. 贪心方法，每一步都找到一个局部影响力最大的节点（也就是说，新增的节点使得增加了最多的活跃节点）

这个方法十分简单，但是计算量庞大，不适合大型网络。

2. 为了降低计算的复杂度，可以采取启发式+贪心的混合算法。

- 首先根据网络的特点，比如节点的出度，节点的影响力选择M个点进行传播。
- 再用贪心算法选择剩下的K-M个节点。

基于社区的影响力最大化

基本思路：由于大型社交网络的计算复杂度太大，就算运用启发式+贪心的混合算法还是比较复杂，将网络分成若干个社区，然后在每个独立的社区上应用已有的影响力最大化算法求解。

可能存在的问题及解决思路：

缺失值填充的方法

1. 如果缺失的feature过多，就直接删除数据。
2. 如果该用户有配偶，则用热卡填充的方式，即使用配偶的数据进行填充。
3. 如果只是缺失少量的feature，因为有一些用户的标签数据（利用微博的数据），可以用K个相似用户的均值来填充。这样就不会和家庭以及同事关系产生交叉影响了。