社交网络影响力传播研究

背景

互联网和大数据的发展，社交网络影响力的传播研究成为数据挖掘和社交网络分析的热点。

三大研究方向

这篇文章主要综述了影响力传播研究的三大方向，以及每个方向目前已经成熟的各种方法。

其三大方向为：

- 影响力传播模型：即要描述影响力如何在社交网络中传播，及其传播的特点和性质。
- 影响力传播学习：即通过大数据学习影响力传播模型的参数。
- 影响力传播优化：即要扩大传播影响力或者控制和减弱不希望传播的影响力。

1. 影响力传播模型

用一个有向图$G=(V,E)$来表示社交网络，每个节点$v \in V$代表社交网络中的每个用户，每条边$(u,v) \in E$代表节点u到v的影响力关系。

Notes：

- 节点 u 到 v 有影响力，但节点 v 到 u 不一定有影响力。
- 边可加权重表示影响力强度
- $N^+(v)$： v 的出邻居； $N^-(v)$： v 的入邻居。对 (u,v) 来说， v 是 u 的出邻居， u 是 v 的入邻居。
- 激活：图中每个节点有两种状态：活跃，不活跃。接受对应实体节点变为活跃，这个过程称为节点被激活。

随机模型

随机模型直接的反映了社交网络中影响力传播的不确定性。

随机模型分类为

- 离散时间和连续时间：影响力传播和节点状态在离散时间点（连续时间轴）上发生。
- 递进和非递进：节点状态从不活跃到活跃不会发生改变（来回切换）。

离散时间递进传播模型

1. 独立级联模型

   每条边$(u,v) \in E$都有一个激活概率$p_{uv} \in [0,1]$，表示通过被激活的节点u将节点v激活的概率。

   其过程如下：

   1. $t=0$ 时，随机选取被激活节点集合，这个集合也称为种子集 S_0 .
   2. $t \geq 1$ 时，对上一时刻被激活的节点 $u \in S_{t-1} \backslash S_{t-2} (S_{-1}=\phi)$ ，尝试对它的每个出邻居 $v \in N^+(u)$ 按照概率 p_{uv} 进行一次激活。若 v 被激活，则 $v \in S_t \backslash S_{t-1}$ ，若 v 没有被其它入邻居激活，则继续处于不活跃状态，即 $v \in S_t$ 。传播直到不再有被激活的节点为止。

   影响力传播主要关注影响力延展度：$\sigma(S_0)$， 即被激活节点个数的期望值$\mathbb{E}[|S_\infty|]$(S_∞为传播结束时所有活跃节点的集合)。

2. 线性阈值模型

   每条边$(u,v) \in E$都有一个权重$w(u,v) \in [0,1]$，且$\sum_{u \in N^-(v)} w(u,v) \leq 1$;每个节点$v \in V$都有一个阈值$\theta_v \in [0,1]$。当$t$时刻，$\sum_{u \in N^-(v)} w(u,v) \geq \theta_v \ (u \ is \ active)$时，$v$处于被激活状态。其过程与独立级联模型类似。

其问题在于，一方面，它的阈值是随机选取的，在实际中，阈值选取有随机性，但应该在更窄的范围内选取；另一方面，更窄的阈值会使模型的分析和计算难道显著增加。因此线性阈值模型面临两难问题。

3. 其它传播模型
    - 连续时间模型

      针对节点的激活连续时间内发生的。
    - 传染病模型

      基于人际之间接触网络的。
    - 选举模型

      非递进模型。
    - 博弈论模型

      反映了人际之间的趋同效应和某些产品的网络外部效应。
    - 多实体传播模型

      用于多个实体同时传播的情况。

2.影响力最大化

影响力最大化就是在给定图结构、影响力传播模型及其参数的情况下，选取k个节点作为种子集合S^*，从而使其影响力延展度最大，即$S^* = argmax_{S\subseteq V,|S|=k}\sigma(S)$ 。

影响力最大化问题在经典模型下是图覆盖问题的一种拓展，属于NP难问题，解决其优化的有效方法是近似算法。

基于子模函数的贪心算法

子模性与单调性：

有限集合V的任意子集映射到实值函数$f:2^v\rightarrow\mathbb{R}$，设任意子集$S\subseteq V$，$S$的超集T满足$S\subseteq T\subseteq V$。若$T$以外的任何一个元素$u\in V\backslash T$，都有$f(S\bigcup\{u\})-f(S)\geq f(T\bigcup\{u\})-f(T)$，则集合函数$f$具有子模性；若$f$满足$f(S)\leq f(T)$则集合函数$f$具有单调性。

单调子模函数可以用贪心算法来求得$(1-1/e)$的近似最优解，但在贪心算法下求$\sigma(S)$是很难的。

蒙特卡罗方法：

可用随机模拟方法来模拟影响力传播，从而估算$\sigma(S)$的值，缺点是效率低。

改进1. 启发式算法

针对独立级联模型的PMIA启发式算法。主要思想是针对某一节点影响力的传播转化为在该节点区域的一颗有代表性的最大影响力传播子树上的传播。而该子树可以用Dijkstra在线性时间内完成。

其它继续改进的算法还有IRIE、LDAG、SIMPATH算法等，但他们都缺少理论保证。

改进2. 针对蒙特卡罗方法做出的改进

1. 偷懒估值法

   利用延展度拖延函数的子模性减少对函数估值的次数。
2. 反向蒙特卡罗

   其思想是不从种子节点去模拟估算种子节点的影响力，而是随机选取图上的点，从该节点出发反方向进行蒙特卡罗模拟，得到的集合实际上是最可能影响该节点的集合。这种方法近线性，有$1-1/e-\epsilon$近似比。

其它优化问题

1. 种子集合最小化
2. 利润最大化
3. 影响力传播监控
4. 多实体传播模型下的影响力最大化
5. 网络拓扑的优化
6. 非子模性的影响力优化问题

3. 影响力传播学习

从数据中挖掘用户行为的传播方式和对应的参数，从而为影响力传播建模和优化服务。其基于的数据分为两类，一类是社交网络结构的数据，另一类是用户的某一类型行为的时间序列。

影响力传播学习的主要思想就是相连的两个用户在相近的时间内先后执行了同样的动作，那么认为这是先执行动作的用户对后执行动作的用户的一次成功影响。

传播影响力学习要解决的两个问题：区分同质性和影响力、判断真正影响节点的邻居。

同质性

再相近的时间内，两个月相连用户先后执行了同样的行为，不一定是先执行用户对后执行用户的影响，有可能是他们的兴趣相同，都来自第三个用户。这种现象就是同质性。因此要区分同质性和影响力，这并不容易，洗牌测试方式对判断影响力存在性有一定影响，但还需继续完善。

判断真正影响节点的邻居

1. 最大似然估计

   主要思想是虽然一个节点有可能被多个邻居节点影响，但是如果实际数据中一个节点的动作经常跟随它某一个邻居节点的动作，这说明这个特定节点对它的影响力可能较大。

2. 基于信用分配的频度分析

   基本思想是当一个节点被激活，要判断是它受哪个邻居节点影响时，将部分信誉积分平摊到所有参与的节点中（积分总和为1），可以平分也可以设置权重，当传播完成时，计算节点总的信誉积分和它被激活邻居节点个数的比值，即频度，就可以判断此节点对它邻居节点的影响力。

存在的问题

影响力传播学习

1. 影响力传播学习方面的准确、有效问题
2. 区分影响力和同质性问题

影响力建模

1. 缺乏实际数据的有效验证
2. 传播的动态性和网络的动态性合理结合

影响力优化

1. 应用有效性需要实际检验