研究快报

基于完全级联传播模型的社区影响最大化

冀进朝[1], 韩 笑[2], 王 喆[1]
(1. 吉林大学 计算机科学与技术学院, 长春 130012; 2. 吉林大学 数学学院, 长春 130012)

摘要: 基于社会网络的社区结构特性和网络中个体间的相互影响, 通过引入社区影响最大化的概念, 并根据节点间相互影响强度的动态变化, 提出一种新的影响传播模型: 完全级联传播模型. 利用该传播模型进行社区影响最大化研究, 在安然邮件数据集上对该传播模型和独立级联模型进行实验对比, 结果表明了该模型在社区影响最大化上应用的有效性.

关键词: 社区影响最大化; 传播模型; 感染力
中图分类号: TP391 **文献标识码:** A **文章编号:** 1671-5489(2009)05-1032-03

Community Influence Maximizing Based on Comprehensive Cascade Diffuse Model

JI Jin-chao[1], HAN Xiao[2], WANG Zhe[1]
(1. *College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*;
2. *College of Mathematics, Jilin University, Changchun* 130012, *China*)

Abstract: In consideration of the community structure existing in social network and individual's interaction, we introduced the concept of community influence maximization. Furthermore, the influence probability among nodes may change due to the dynamic change of the interaction's intensity between the nodes. Thus, in this paper, the authors will propose a new diffuse model named comprehensive cascade model and use this model to study community influence maximization. Through the experiments on Enron email dataset, we compared the model's performance with independent cascade's. The result shows that the model is feasible in the community influence maximization.

Key words: community influence maximizing; diffuse model; contagion

在社会网络中, 通常个体用节点表示, 个体之间的联系用边表示. 传统的影响最大化问题[1-2]关心的是在网络中发现最有影响力的 k 个节点作为起始对象, 目的是使最终受到影响(如采纳某项建议、购买某种商品等)的个体数最多. 节点间的边被分配一个介于0和1之间的权值. 这个值表示当其中一个节点被激活时, 它影响另一个节点的概率. 实际上, 社会网络具有社区结构, 即网络中的节点因为某些因素(如朋友、亲戚、共同的业余爱好等)形成了一些社区. 同社区中的节点具有一些共同特征, 社区内的节点联系密切, 社区间的联系松散. 目前, 人们已提出了许多社区识别算法[3-5]. Kempe等人[6-7]曾提出一个考虑节点间影响衰减的递减级联传播模型. 事实上, 节点间影响不仅存在着衰减, 也存在着加强和维持原状. 基于这种思想, 本文提出一个新的影响传播模型, 并利用该模型研究社区

收稿日期: 2009-05-31.

作者简介: 冀进朝(1982~), 男, 汉族, 硕士研究生, 从事数据挖掘的研究, E-mail: jinchao0374@163.com. 通讯作者: 韩 笑(1980~), 女, 汉族, 博士, 讲师, 从事计算机图像图形学、模式识别和微分方程图像处理的研究, E-mail: hanx@jlu.edu.cn.

基金项目: 国家自然科学基金(批准号: 60673099; 60873146)、国家高技术研究发展计划863项目基金(批准号: 2007AA04Z114; 2009AA02Z307).

影响最大化.

1　完全级联传播模型

定义 1　节点所具有的影响其他节点的能力称为感染力. 具备感染力的节点能够影响其他节点；不具备感染力的节点不能影响其他节点.

定义 2　设 V 表示社会网络的个体集，E 是边集(边表示节点间的联系)，则社会网络图记为 $G(V,E)$. 如果两个节点之间有联系，则它们之间存在一条边；反之，节点间不存在边.

定义 3　对于网络图 $G(V,E)$ 中的一个节点 v，与 v 有边直接相连的节点称为节点 v 的邻居节点.

定义 4　节点 v 所有邻居节点构成的集合称为邻居节点集 $N(v)$.

定义 5　如果一个社区中有一个节点被激活，则称这个社区被覆盖了.

通常，影响最大化的关键是在网络中发现最有影响力的 k 个节点. 将社区影响最大化问题变为选择最好的 k 个节点初始激活，目的是在影响最大化过程的最终阶段使得社区覆盖最大. 文献[6-7]介绍了描述节点激活行为的几种模型，其中线性阈值模型和独立级联模型是两种最基本的传播模型. 在考虑社会网络中诸如某种观念、某项创新等影响的传播模型时，可把一个节点的状态表示为活跃和不活跃两种情况，且一个节点只能从不活跃状态转变为活跃状态. 如果一个节点有越来越多的邻居节点变活跃，则它也越来越趋向于活跃. 因此传播过程可以从一个初始不活跃节点 v 的角度看作如下过程：随着时间的推移，v 的邻居节点中有越来越多的节点变活跃；在某个时间点上，这可能使 v 变活跃，并且 v 的决策可能会依次触发与 v 相连节点的决策. 当一个节点在时间步 t 首先变活跃时，认为它具有感染力，它具有影响每个不活跃邻居节点 v 的一次机会. 一次成功的激活尝试将使 v 在下一个时间步 $t+1$ 成为活跃节点. 如果 v 的多个邻居节点在时间步 t 变活跃，则这些活跃的邻居节点按任意顺序尝试激活节点 v，但所有的这些尝试都发生在时间步 t. 一个活跃节点 u 对其所有邻居节点尝试激活后，仍保持活跃，但已不具备感染力了. 当不存在具备感染力的节点时，这个过程结束. 为了充分描述这个模型，需要给出节点 u 尝试激活节点 v 的成功概率. 在最简单的独立级联模型中，成功激活概率是一个与传播过程无关的常量 $p_v(u)$. 而在现实生活中存在着个体间影响概率动态变化的情况，且这种变化与之前的交互有关. 因此，本文提出一个新的影响传播模型：完全级联传播模型. 在该模型中活跃节点 u 影响节点 v 的概率是节点 v 的邻居中已经试图激活 v 但未激活成功节点集的一个可增减函数. 如果用 S 表示 v 的邻居中已经尝试激活 v 但未激活成功的节点集，则活跃节点 u 影响不活跃的邻居节点 v 的概率为

$$p_v(u,S) = p_v(u) - k \times \frac{|S|}{|V|} p_v(u),$$

其中 k 是从 $\{-1,0,1\}$ 中随机选择的一个值，$|V|$ 是整个网络图的节点个数，$S \subseteq N(v)$，$|S|$ 是 S 中的节点个数，$p_v(u)$ 是节点 u 影响节点 v 的初始概率. 因为缺乏节点间影响何时该增强、何时该减弱的相关知识，所以用随机选择方法模拟这种情况. 该模型不仅表示了每个节点如何影响其他节点，而且表示了节点的影响力如何受节点之前的交互所影响. 具有感染力的节点是以动态变化的概率激活其邻居节点的.

为了简化问题，假设完全级联传播模型具有顺序无关性. 假设 T 是节点 v 的活跃邻居节点集，显然，$T \subseteq N(v)$. 顺序无关性的含义是：如果集合 T 中的所有节点试图影响 v，则他们尝试激活 v 的顺序不影响节点 v 最终成为活跃节点的概率. 用公式表示为：如果 $u_1,\cdots,u_r$ 和 $u_1',\cdots,u_r'$ 是 T 的两个排列，并且 $T_i=(u_1,\cdots,u_{i-1})$，$T_i=(u_1',\cdots,u_{i-1}')$，顺序无关性意味着：

$$\prod_{i=1}^{r}(1-p_v(u_i,S \cup T_i)) = \prod_{i=1}^{r}(1-p_v(u_i',S \cup T_i')),$$

对所有与 T 不相交的集合 S 都成立.

2　实验结果分析

本文假设每个社区至少有两个节点，一个网络至少有两个社区. 采用文献[5]中的 Radicchi 算法进行社区划分. 在安然邮件数据集(Enron email dataset，http://www.cs.cmu.edu)上进行测试，通过在

相互交流超过5封电子邮件的管理人员之间建立一条链接关系形成社会网络. 对两种不同算法在独立级联传播模型和完全级联传播模型上得到的结果进行实验对比, 比较覆盖的社区数. 随机方法随机选择 k 个节点, 度方法选择度最高的 k 个节点. 在社区影响最大化的实验中, 把所有边的权值初始化为0.01, 然后在完全级联模型和独立级联模型上分别实验. 由于模型的随机性, 对每种方法重复模拟1 000次, 然后求平均结果. 对于安然邮件数据集, 设初始活跃节点为10个节点, 实验结果列于表1.

表1 两种传播模型上社区覆盖数的比较

Table 1 Comparison between community coverages on two diffuse models

算法	社区覆盖数	
	独立级联传播模型	完全级联传播模型
随机方法	5.3	5.4
度方法	6.9	6.8

由表1可见, 完全级联传播模型的社区覆盖面大致等于独立级联传播模型的社区覆盖面, 这是因为在完全级联模型中考虑了影响的动态变化, 证明了完全级联传播模型在社区影响最大化上的有效性和可行性.

综上可见, 社区影响最大化的研究离不开合适的影响传播模型. 由于节点交互的动态性, 节点间影响概率也是动态变化的. 因此, 本文提出一个新的影响传播模型, 并利用其进行社区影响最大化研究. 通过在安然邮件数据集上对该传播模型和独立级联传播模型进行实验对比, 表明了此模型在社区影响最大化上的有效性.

参 考 文 献

[1] Domingos P, Richardon M. Mining the Network Value of Customers [C]//Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. California: ACM, 2001: 57-66.

[2] Richardson M, Domingos P. Mining Knowledge-sharing Sites for Viral Marketing [C]//Proceedings of the Eighth Intl Conf on Knowledge Discovery and Data Mining. Alberta: ACM, 2002: 61-67.

[3] Radicchi F, Castellano C, Cecconi F, et al. Defining and Identifying Communities in Networks [J]. PNAS, 2004, 101(9): 2658-2663.

[4] CAI Hua, ZHOU Chun-guang, LU Ting-yu, et al. OCSMA: an Alogorithm to Mine Overlapping Community Structure in Networks [J]. Journal of Jilin University: Engineering and Technology Edition, 2009, 39(4): 1035-1040. (才 华, 周春光, 卢廷玉, 等. 重叠社区结构的挖掘算法 [J]. 吉林大学学报: 工学版, 2009, 39(4): 1035-1040.)

[5] ZHOU Chun-guang, QU Peng-cheng, WANG Xi, et al. DSNE: a New Dynamic Social Network Analysis Algorithm [J]. Journal of Jilin University: Engineering and Technology Edition, 2008, 38(2): 408-413. (周春光, 曲鹏程, 王 曦, 等. DSNE: 一个新的动态社会网络分析算法 [J]. 吉林大学学报: 工学版, 2008, 38(2): 408-413.)

[6] Kempe D, Kleinberg J, Tardos É. Maximizing the Spread of Influence through a Social Network [C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington: ACM, 2003: 137-146.

[7] Kempe D, Kleinberg J, Tardos É. Influential Nodes in a Diffusion Model for Social Networks [M]. Berlin: Springer, 2005, 3580: 1127-1138.