移动社会网络信息传播模型构建与 top-k 节点挖掘

史文国[1]，王 瑜[2]
（1. 中国科学院电子学研究所，北京 100190；2. 北京大学 信息科学技术学院 智能科学系，北京 100871）

摘 要：在移动社会网络中挖掘出有影响力的 top-k 节点，对于移动运营商作出新产品或服务战略营销决策至关重要。针对移动社会网络的特点，提出一种充分考虑移动社会网络特点的信息传播模型以及基于该模型的 top-k 节点挖掘算法。实验证明，该方法能准确高效地定位移动社会网络中的活跃节点，这对于移动运营商作出营销决策起着至关重要的作用。

关键词：信息传播模型；移动社会网络；top-k 节点挖掘算法
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2012)08-2830-03
doi:10.3969/j.issn.1001-3695.2012.08.007

Diffusion model and top-k nodes mining for mobile social networks

SHI Wen-guo[1], WANG Yu[2]
(1. *Institute of Electronics, Chinese Academy of Sciences, Beijing* 100190, *China*; 2. *Dept. of Machine Intelligence, School of Electronics Engineering & Computer Science, Peking University, Beijing* 100871, *China*)

Abstract: The top-k nodes in those network systems often play critical roles in information exchanges and spreading. Thus finding those influential carriers in mobile social networks is very useful for mobile operators to make strategy, such as sales marketing etc. This paper proposed a top-k mining algorithm based on the information diffusion model. The experiments prove that the proposed algorithm can mine influential nodes efficiently and accurately in mobile social networks.

Key words: information diffusion model; mobile social networks; top-k nodes mining algorithm

近些年，社会网络分析（SNA）越来越引起各个领域的学者的关注，在社会网络中进行市场营销也成为各商家和运营商探索的热点。传统的市场营销就是将产品全部推销给潜在顾客。而病毒式营销则是首先选择那些最有希望接受该产品的顾客，然后再由这些顾客将产品进一步推销出去，这就节省了营销成本；并且如果营销对象选择得恰当的话，可以用最少的成本换取最大化的营销效益，一举两得，这就是病毒式营销的魅力所在。

移动社会网络是由移动通话数据抽象出来的社会网络，每个手机号都是网络节点，手机号码间的联系构成网络的边。准确全面地了解移动社会网络，对于移动运营商进行新产品和服务的推广具有深远而重大的意义。所以，在移动社会网络中挖掘具有影响力的 top-k 节点，对于移动运营商作出营销决策、获取营销效益的最大化具有十分重要的意义。

近几年，在社会网络中利用数据挖掘算法挖掘该网络中具有影响力的 top-k 节点已成为学术界研究的热点难点问题[1~3]。Domingos 等人[4]提出解决该问题的基本算法，即将该问题抽象为一个信息传播的概率模型来解决，并且选择基于该概率模型在网络上传播信息最大节点集合作为活跃节点集合，这成为该问题研究的先驱。但该问题是 NP 难问题，且基本算法由于节点间传播范围的重叠性无法保证信息的传播精度，研究学者们用两种思路解决该问题。一些学者[5,6]的工作给出了该问题更精确的贪心近似算法，并证明该算法是(1 - 1/e)近似于最优解的；另外一些对于应用感兴趣的学者[7]试图提出其他信息传播模型和数据挖掘算法来解决在社会网络中 top-k 节点挖掘问题。所有这些研究者忽视的一个问题就是理论和应用结合，他们并没有提出一个实用的信息传播模型并在真实的网络数据上进行证明。

本文提出一个基于移动社会网络的信息传播模型，并利用该信息传播模型进行市场营销的信息传播。

1 相关工作

下面介绍基本的信息传播模型和 top-k 活跃节点挖掘算法。SIS(susceptible-infective-susceptible)模型在文献[8]中提出，是传染病传播的最经典模型。它用来估计某种传染病在人群中的传播范围，它需要根据疾病传播能力的不同给出一个传播速度 v，传播就可以表示为传播速度、节点度数和被传播节点个数的函数，该模型也可以应用于信息传播中。文献[5]提出了递减的瀑布信息传播模型(decreasing cascade model)，即当一个节点首次被信息传播到后，它的状态转换为活跃—传染，而一个活跃—传染状态的节点会以一定的概率去激活它的处于非活跃状态的邻居节点。文献[6]提出了一个线性阈值模型(linear threshold model)，其基本思想就是一个节点转换为活跃状态的概率会随着它周围活跃节点个数的增多而增大，如果一个节点的邻居中活跃节点的数量超过一定的阈值，该节点将转换为活跃状态。

简单 top-k 活跃节点挖掘算法就是计算给定的一个社会网络 $G(V,E)$ 中每一个节点的传播范围，并且选择传播范围最大的前 k 个节点组成 top-k 节点集合[7]。这是解决 top-k 问题最原

收稿日期：2011-12-30；**修回日期**：2012-02-16

作者简介：史文国(1986-)，男，山东烟台人，硕士研究生，主要研究方向为智能信息系统、信号处理(flame12334@163.com)；王瑜(1987-)，女，河北秦皇岛人，硕士研究生，主要研究方向为人工智能、数据挖掘、社会网络分析.

始的思想,但是这个近似算法无法保证结果的精确度。

2　基于信息传播模型的 top-k 节点挖掘

在此部分,首先提出一个基于移动社会网络的信息传播模型,再利用贪心算法基于此信息传播模型挖掘 top-k 节点。

给定移动社会网络和挖掘活跃节点的个数 k,再在信息传播模型上用贪心算法挖掘 top-k 活跃节点集合。

2.1　基于交互式马尔科夫链的信息传播模型

移动社会网络,就是利用移动用户的手机通话数据构建的社会网络。在移动社会网络中,提出基于交互式马尔科夫链的信息传播模型[9],该模型根据移动社会网络的特性构建,仅适用于移动社会网络的信息传播规律。考虑通话次数、通话双方等信息,移动社会网络是一个有向加权网络。

2.1.1　信息传播模型基本思想

在本信息传播模型中,网络中的节点有三种状态:活跃—传染、活跃—非传染、非活跃。而节点间的状态转换遵循以下规则:a)当一个活跃—传染节点 i 与一个非活跃节点 j 接触时,节点 j 转换为活跃—传染状态的概率为 γ,该规则描述了一个人接受某种产品的概率,并且状态转换的概率随着通话次数的增加而增大,即通话次数越多,传播越容易发生;b)当一个活跃—传染节点 i 与另一个处于活跃—传染状态或者活跃—非传染状态的节点 j 通话后,节点 i 转换为活跃—非传染状态的概率为 α,该规则描述了当一个人通过跟他人交流充分了解了某产品后,对此产品失去兴趣不再关注的概率。

A 表示移动社会网络中初始的活跃—传染状态节点集合,即进行市场营销首先进行产品或服务投放的顾客,在时刻 $t=0$ 即为活跃状态。节点 u 和 v 表示网络中的两个节点,$sp(u,v)$ 表示节点 u 和 v 间的网络最短路径距离,$sp(A,v)$ 表示集合 A 到节点 v 的最短距离,定义为

$$sp(A,v)=\min_{u\in A}sp(u,v) \tag{1}$$

如果集合 A 中没有节点可达 v,则 $sp(A,v)=\infty$。

同理,用 $lp(u,v)$ 表示节点 u 和 v 间的网络最长路径距离,$lp(A,v)$ 表示集合 A 到节点 v 的最长距离,有

$$p(A,v)=\min_{u\in A}lp(u,v) \tag{2}$$

2.1.2　节点状态转换概率定义

定义 1　节点保持非活跃状态的概率。节点 j 在时刻 t 处于非活跃状态,用 pro_j^{ii} 表示该节点在时间区间$[t,t+\Delta t]$内保持非活跃状态的概率,而 pro_j^{ic} 表示节点 j 在此时间区间内转换为活跃—传染状态的概率,则有

$$\mathrm{pro}_j^{ii}(t)=\begin{cases}1 & \gamma t<sp(A,j)\\ (1-\gamma\Delta t)^c & sp(A,j)\leqslant\gamma t\leqslant lp(A,j)\\ 1 & \gamma t>lp(A,j)\end{cases} \tag{3}$$

其中:$c=c(t)$ 表示指向节点 j 的节点中在时刻 t 处于活跃—传染状态的数量。式(3)可以解释为在时间区间$[t,\ t+\Delta t]$内,所有指向节点 j 的活跃—传染状态的节点都未能成功将信息传播给 j,而且 j 的状态仅会在 $sp(A,j)\leqslant\gamma t\leqslant lp(A,j)$ 时发生改变,若超出该范围,信息无法到达节点 j 进行传播。

定义 2　引入网络权重的传播概率。在移动社会网络中,令通话次数作为边的权重,此时信息传播概率定义为

$$\gamma_{ij}=2\bar{\gamma}\frac{w_{ij}}{w_{\max}+w_{\min}} \tag{4}$$

其中:γ_{ij}表示节点 i 传播给节点 j 的概率;$\bar{\gamma}$ 是该信息的平均传播概率;w_{ij}是边 e_{ij}的权重;$w_{\max}$和 $w_{\min}$分别为网络边的最大权重和最小权重。

定义 3　指向节点的活跃—传染状态邻居分布。若有 k 个节点指向节点 j,每个节点处于何种状态在整个网络中可近似认为是独立的,那么 $c(t)$可以视做一个满足二项分布的随机变量:

$$H(c,t)=C_k^c\varphi(k,t)^c(1-\varphi(k,t))^{k-c} \tag{5}$$

其中:$\varphi(k,t)$是指向度数为 k 的节点中在 t 时刻处于活跃—传染状态的概率,故将 $\varphi(k,t)$表示为

$$\varphi(k,t)=\sum_{e_{ij}\in E}q(i,k)\,\mathrm{pro}_i^s(t) \tag{6}$$

$$q(i,k)=\frac{k'P(k')+kP(k)}{2\bar{k}} \tag{7}$$

其中:$q(i,k)$是度数为 k 的节点 j 与度数为 k'的节点 $i(e_{ij}\in E)$的关联函数;pro_i^s 是 t 时刻节点 i 处于活跃—传染状态的概率;$\bar{k}$ 是整个网络的平均度数。

综上所述,将分析结果代入式(3)可得

$$\mathrm{pro}_j^{ii}(k,t)=\begin{cases}1 & \gamma t<sp(A,j)\\ (1-\gamma\Delta t)\sum_{e_{ij}\in E}q(i,k)\,\mathrm{pro}_i^s(t))^k & sp(A,j)\leqslant\gamma t\leqslant lp(A,j)\\ 1 & \gamma t>lp(A,j)\end{cases} \tag{8}$$

同理,可以得到一个度数为 k 的节点在时间区间$[t,t+\Delta t]$内,保持活跃—传染状态的概率:

$$\mathrm{pro}_j^{ss}(k,t)=\begin{cases}1 & \gamma t<sp(A,j)\\ (1-\alpha\Delta t)\sum_{e_{ij}\in E}q(i,k)(\mathrm{pro}_i^s(t)+\mathrm{pro}_i^i(t))^k & sp(A,j)\leqslant\gamma t\leqslant lp(A,j)\\ 1 & \gamma t>lp(A,j)\end{cases} \tag{9}$$

还可得到用来刻画一个度数为 k 的节点在时间区间$[t,t+\Delta t]$内由非活跃状态转换为活跃—传染状态的概率。

2.1.3　各状态节点个数期望分析

用 $I(k,t)$、$S(k,t)$、$N(k,t)$分别表示移动社会网络中处于非活跃、活跃—传染、活跃—非传染状态的概率,一个非活跃状态的节点在时间区间$[t,\ t+\Delta t]$转换为活跃—传染状态的概率$(1-p_j^{ii})$符合泊松分布,时间区间后整个网络中由非活跃转换为活跃—传染状态的个数期望为 $I(k,t)(1-\mathrm{pro}^{ij})$。由此可得网络中非活跃状态的节点数量期望为

$$I(k,t+\Delta t)=I(k,t)-I(k,t)\times(1-(1-\gamma\Delta t)\sum_{e_{ij}\in E}q(i,k)\,\mathrm{pro}_i^s(t))^k \tag{10}$$

同理,可以得到 $S(k,t)$、$N(k,t)$的数量期望分别为

$$S(k,t+\Delta t)=S(k,t)+I(k,t)(1-(1-\gamma\Delta t\sum_{e_{ij}\in E}q(i,k)\,\mathrm{pro}_i^s(t))^k)-S(k,t)(1-(1-\alpha\Delta t\sum_{e_{ij}\in E}q(i,k)(\mathrm{pro}_i^s(t)+\mathrm{pro}_i^i(t)))^k) \tag{11}$$

$$N(k,t+\Delta t)=S(k,t)+S(k,t)\times(1-(1-\alpha\Delta t\sum_{e_{ij}\in E}q(i,k)(\mathrm{pro}_i^s(t)+\mathrm{pro}_i^i(t)))^k) \tag{12}$$

令 $\rho^x(k,t)$表示度数为 k 的节点在时刻 t 处于状态 x 的密度,网络中节点仅有三种状态,有 $\rho^i(k,t)+\rho^s(k,t)+\rho^n(k,t)=1$。在以上三式中令 $\Delta t\to 0$,可得

$$\frac{\partial\rho^i(k,t)}{\partial t}=-k\gamma\rho^i(k,t)\sum_{u\in V}q(u,k)\,\mathrm{pro}_u^s(t) \tag{13}$$

$$\frac{\partial\rho^s(k,t)}{\partial t}=k\gamma\rho^i(k,t)\sum_{u\in V}q(u,k)\,\mathrm{pro}_u^s(t)-k\alpha\rho^c(k,t)\sum_{u\in V}q(u,k)(\mathrm{pro}_u^s(t)+\mathrm{pro}_u^n(t)) \tag{14}$$

$$\frac{\partial\rho^n(k,t)}{\partial t}=k\alpha\rho^c(k,t)\sum_{u\in V}q(u,k)(\mathrm{pro}_u^s(t)+\mathrm{pro}_u^n(t)) \tag{15}$$

通过建立以上模型,可以在营销推广时用以上信息模型进行信息传播模拟,进而挖掘网络中的 top-k 活跃节点。当$t\to\infty$时,$\rho^s(k)=0$,$\rho^n(k)$是最终获取信息的人的密度,本文的任务

就是找到 k 个节点,使得最终的 $\rho^n(k)$ 最大化。

2.1.4 移动社会网络中的传播期望

利用以上信息传播模型令 $t\to\infty$ 解方程组(10)~(12),设传播范围函数 $\bar{R}=\sum_k\rho^n(k,\infty)$,可得

$$R=\sum_k P(k)(1-e^{-\gamma k w_\infty}) \tag{16}$$

$$w_\infty=\frac{2(\gamma-\gamma_c)}{\gamma^2\sum_k q(k)k^2(\gamma_c+2\alpha)} \tag{17}$$

其中:γ_c 是传播阈值,当 $\gamma\geqslant\gamma_c$ 时,信息传播可以发生;当 $\gamma<\gamma_c$ 时,信息传播将不会发生。在移动社会网络中,γ_c 为

$$\gamma_c=\frac{1}{\sum_k(q(k)k)} \tag{18}$$

在移动社会网络中,用 degree(k) 表示网络中度数为 k 的节点的个数,度数关联函数 $P(k)$ 可以表示为

$$P(k)=\frac{\text{degree}(k)}{\sum_j \text{degree}(j)} \tag{19}$$

2.2 top-k 活跃节点挖掘算法

本文的优化目标是信息传播模型中的传播范围 R 最大化,故最终使得传播范围 $N=\{R(1)\cup R(2)\cup\cdots\cup R(k)\}$ 最大化。简单 top-k 活跃节点挖掘算法忽视了网络中节点传播范围的重叠性,而贪心算法解决了该问题,算法描述如下所示。

```
algorithm 1:greedy top-k algorithm
input:mobile social network G(V,E),information diffusion model R.
output:top-k nodes set S.
a)for each node i do
b)    state(i) = active;state(j) = inactive (j∈V and j≠i);
c)    compute R(i);
d)end
e)U = {R(1),R(2),…,R(n)}, N = ∅;
f)for count = 1 to k do
g)    select R(i) which maximizes {R(i) - N∩R(i)};
h)    N = N∪R(i);U = U - R(i)
i)    output node i;
j)end;
```

从第 a)~d)步,依次计算每个节点的传播范围,而第 g)步选择产生传播范围增益最大的节点作为 top-k 活跃节点。

3 实验

3.1 数据集

自从手机广泛应用以来,广大手机用户每时每刻的通话行为产生大量的通话数据,而信息也随着这些通话行为广泛地传播。本文利用这些数据构建移动社会网络,进而研究移动社会网络上的信息传播问题。

数据来源是中国某省份 2007 年 7~9 月的通话记录(CDR),采用随机的方式从中抽取 723 201 个用户的通话记录构建社会网络,网络的平均度数为 13.4,度数分布情况如图 1 所示呈现幂率分布,表示移动社会网络是一个尺度无关网络。

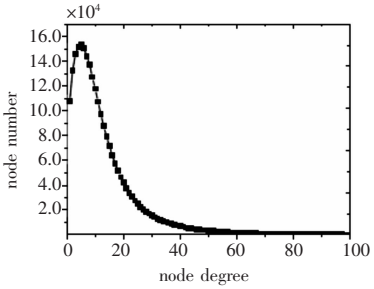

图 1 移动社会网络度数分布

3.2 单个节点的信息传播

图 2 展示了从网络中随机抽取单个节点的传播范围随传播概率 γ 的变化情况,其中恢复概率 $\alpha=\{0.00,0.25,0.50,0.75,1.00\}$,采用式(18)计算。

从图 2 中可以看出,当 $\gamma>\gamma_c(0.089\ 9)$ 时,信息传播才得以进行,随着传播概率 γ 的增大,传播范围也随之增大;而当传播概率相同时,传播范围随着恢复概率 α 的增大而减小。

3.3 挖掘 top-k 活跃节点

从图 2 可以看出,α 限制了信息的传播,但是并未改变信息的传播趋势,故 α 取值并不影响对信息传播趋势的研究。在此部分,令 $\alpha=0$,并设定传播概率 $\gamma=\{0.2,0.3,0.4,0.5\}$,采用贪心算法挖掘 top-k 活跃节点,结果如图 3 所示。

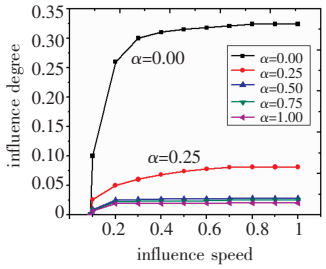

图 2 单个节点信息传播范围

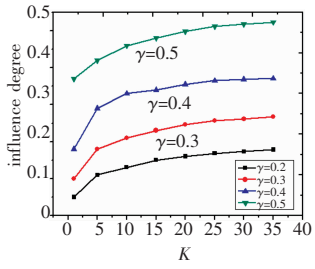

图 3 贪心算法挖掘 top-k 活跃节点的信息传播范围随 k 的变化

从图 3 中可以看出,随着传播源节点个数 k 的增加,最终的信息传播范围也会随之增加;其次,随着传播概率的增大,最终传播范围也随之相应增大。

从实验结果中可以看出,本文提出的信息传播模型可以符合常理地模拟移动社会网络中的信息传播过程,而且该模型挖掘的活跃节点也比较均匀地分布在社会网络中(图 4、5)。这说明本模型对商家进行产品推广和广告投放有很现实的意义。

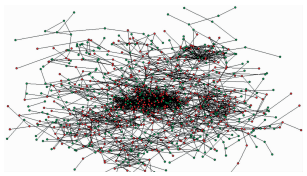

图 4 γ=0.5 时信息传播范围

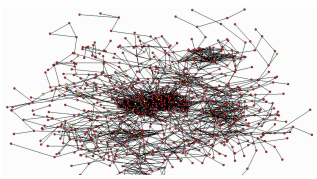

图 5 top-k 活跃节点

(电子版图 4、5 中,红色代表活跃非传染状态;绿色代表非活跃状态;黄色代表活跃节点)

4 结束语

本文提出一种基于移动社会网络市场营销和广告投放的信息传播模型,并且利用常规贪心算法在该信息传播模型上进行 top-k 活跃节点挖掘实验,该信息传播模型针对移动社会网络的社交规律来刻画移动社会网络的信息传播,是对移动社会网络中信息传播规律描述的新尝试。

参考文献:

[1] CHEN Wei, WANG Chi, WANG Ya-jun. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]// Proc of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 1029-1038.

[2] CHEN Wei, WANG Ya-jun, YANG Si-yu. Efficient influence maximization in social networks[C]//Proc of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 199-208.

[3] WANG Yu, CONG Gao, SONG Guo-jie, *et al*. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks[C]//Proc of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 1039-1048.

(下转第 2844 页)

(上接第 2832 页)

[4] DOMINGOS P, RICHARDSON M. Mining the network value of customers[C]//Proc of the 7th International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2001: 57-66.

[5] KEMPE D, KLEINBERG J, TARDOS E. Influential nodes in a diffusion model for social networks[C]//Proc of the 32nd International Conference on Automata, Languages and Programming. Berlin: Springer-Verlag, 2005: 1127-1138.

[6] KEMPE D, KLEINBERG J, TARDOS E. Maximizing the spread of influence through a social network[C]//Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 137-146.

[7] MA Hao, YANG Hai-xuan, LYU M, *et al*. Mining social networks using heat diffusion processes for marketing candidates selection[C]//Proc of the 17th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2008: 233-242.

[8] SURI R N, NARAHARI Y. Determining the top-k nodes in social networks using the shapley value[C]//Proc of the 7th International Conference on Autonomous Agents and Multiagent Systems. 2008: 1509-1512.

[9] NEKOVEE M, MORENO Y, BIANCONI G, *et al*. Theory of rumour spreading in complex social networks[J]. Physica A: Statistic Mechanics and Its Applications, 2007, 374(1): 457-470.

[10] LÓPEZ-PINTADO D. Diffusion in complex social networks[J]. Games and Economic Behavior, 2008, 62(2): 573-590.

[11] ASAVTHIRATAM C, ROY S, LESIEUTRE B, *et al*. The influence model[J]. IEEE Control Systems, 2001, 21(6): 52-64.

[12] KIMURA M, SAITO K. Tractable models for information diffusion in social networks[C]//Proc of the 10th European Conference on Principles and Practice of Knowledge Discovery in Database. Berlin: Springer-Verlag, 2006: 259-271.

[13] TONG Chang. Analysis of some popular mobile social network systems [C]//TKK T-110.5190 Seminar on Internetworking. 2008.