RAISE:一种高效的社交网络影响成本最小化算法

孙永樾 李红燕 张金波

(北京大学信息科学技术学院 北京 100871) (北京大学机器感知与智能教育部重点实验室 北京 100871)

摘 要 在市场营销、政治选举等领域,说服个体接受新产品或新思想需要耗费一定的成本。将影响成本最小化问题定义为如何选择不同个体,使影响最终扩散到社交网络中给定数量的个体,且耗费的成本最小。运用现有方法解决该问题,解的质量和时间效率都面临一定的瓶颈。为了解决该问题,提出了一种高效的算法——RAISE 算法。在理论上,当期望达到的影响与网络规模可比拟时,该算法具备常数近似比和线性时间复杂度。实践表明,该算法在解的质量和时间效率两方面都显著优于现有方法。

关键词 成本,影响成本最小化,随机采样,在线社交网络

中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.09.007

RAISE:Efficient Influence Cost Minimizing Algorithm in Social Network

SUN Yong-yue LI Hong-yan ZHANG Jin-bo

(School of Electronics Engineering and Computer Science,Peking University,Beijing 100871,China)

(Key Laboratory of Machine Perception(Ministry of Education),Peking University,Beijing 100871,China)

Abstract In many scenarios,such as viral marketing and political campaign,persuading individuals to accept new products or ideas requires a certain cost. Influence cost minimization problem is defined as choosing an influential set of individuals so that the influence can be spread to given number of individuals while the total cost is minimized. The solution quality and time efficiency are faced with bottlenecks when solving this problem with existing methods. To tackle the issue,this paper proposed an efficient algorithm,RAISE. In theory,when the expected influence is comparable to the network size,the proposed algorithm has constant approximation ratio and linear time complexity. In practice,the proposed algorithm is significantly superior to the existing methods in terms of solution quality and time efficiency.

Keywords Cost,Influence cost minimization,Random sampling,Online social network

1 引言

随着 Web2.0 时代的到来,社交网络在诸多领域都展现出了巨大的应用价值,例如市场营销[1]、流言控制[2]、推荐系统[3]等领域。在社交网络的信息传播过程中,成本起到了至关重要的作用。因此,成本的最优化问题在这些领域中得到了越来越多的关注与应用。例如,竞选活动中候选人常利用社交网络来扩大影响。图 1 给出了一个以美国大选为背景的社交网络。

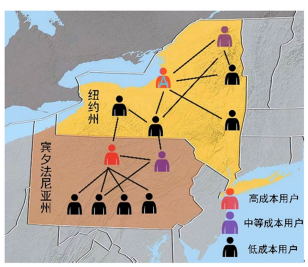

图 1 美国大选示例

Fig. 1 Example of general erection of US

在大选中,候选人为了在一个州内获胜,需要在该州获得至少 50%的支持率。为节约竞选资金,候选人自然会关心如何以成本最小的方式来赢得选举。类似的需求在市场营销、流言控制等领域中也广泛存在。

目前已有许多研究关注成本的最优化问题,但它们普遍面临以下问题。

1)缺乏令人满意的近似比保证,在不同结构网络中的表现差异较大。现有方法具备 $\ln(n)$ 的近似比(n 为网络中节点的数量),即在最差情况下,现有方法返回的可行解的成本是最优成本的 $\ln(n)$ 倍。在大数据时代,节点数以亿计地存在于社交网络中,$\ln(n)$ 约为 20,这意味着现有方法可能造成严重的资金浪费。现有方法在最差情况下表现为网络中节点的度中心性服从特定的幂律分布。以图 1 所示的美国大选为例,宾夕法尼亚州社交网络的节点的度中心性近似接近最差情况。而在纽约州社交网络中,节点的度中心性分布偏离最差情况。由图 1 可知,相同算法在不同州的表现差异巨大,这说明现有方法的解的质量强烈依赖于网络结构。

2)运行时间普遍较长,难以满足即时查询的需求。本文

到稿日期:2018-07-02 返修日期:2018-09-11

孙永樾(1996—),男,硕士生,主要研究方向为数据挖掘,E-mail:redhated@163.com;李红燕(1970—),女,博士,教授,主要研究方向为数据管理与数据挖掘,E-mail:leehy@pku.edu.cn(通信作者);张金波(1992—),男,博士生,主要研究方向为数据管理与数据挖掘。

证明了基于蒙特卡洛模拟方法的算法的时间复杂度为 $O(Jmn \cdot MCtimes)$，其中 J 为期望达到的影响，n 和 m 分别为网络的节点数和边数，$MCtimes$ 代表估计一个集合的期望影响所需要的蒙特卡洛模拟次数。对于节点数量在数万级别的社交网络，当 $MCtimes$ 取值为 10 000 时，这种算法需要数个小时乃至几天的时间来返回一个可行解。在现实情况下，社交网络不但规模巨大，而且频繁地动态变化，现有的基于蒙特卡洛模拟方法的算法很难满足政治选举、市场营销等情境中普遍的即时查询需求。

3)忽略了不同个体具备不同成本的事实，缺乏对成本的合理度量。在现实情境下，说服一个个体接受一种新观点或使用一种新产品的困难程度与个体的特征高度相关，如个体的性别、年龄、兴趣、角色、地理位置等。因此，不同个体的成本一般不同。以图 1 所示的美国大选为例，在纽约州，个体 A 的影响力很大，但是说服他所需的成本却相当昂贵，反而不是较好的选择。然而，个体的成本通常难以获取，也缺乏合理的度量方式。因此，现有的研究普遍简化了这样的情境，认为不同个体的成本是相同的。

针对以上问题和挑战，本文主要做了以下几方面的工作：1)对现有研究进行泛化，形式化定义了影响成本最小化问题，并证明了其难解性；2)灵活地度量了成本，提出了一种有效的成本计算方式；3)提出了基于随机采样的 RAISE 算法，并给出了运行时间和近似比的保证；4)通过实验验证了 RAISE 算法在解的质量和运行时间等方面明显优于已有算法。

2 相关工作

在影响成本最小化问题研究领域中，Chen[4] 率先提出目标集选择问题：寻找一个基数尽可能小的初始集合，使其影响的扩散能够超过给定阈值。该问题忽略了节点之间的成本差异，是影响成本最小化问题的一种特殊情况。之后的研究者在不同的信息传播模型中讨论了影响成本最小化问题。根据解决方案的不同，这些研究大致分为两类：基于对偶问题的近似算法和基于其他方案的近似算法。

基于对偶问题的近似算法中，比较有代表性的是 Long 等[5]提出的 J-Min-Seed Greedy 算法和 Zhang 等[6]提出的 MinSeed-PCG 算法。影响成本最小化问题的目标函数在常见的信息传播模型中不具备单调性和次模性[5-6]，难以直接优化。而其对偶问题——影响最大化问题的目标函数具备单调性和次模性[7]，易于进行优化。基于对偶问题的近似算法利用了这一性质，每次选择边际收益最大的节点，直至满足要求。这类方法通常利用蒙特卡洛模拟方法来估计给定节点集合的期望影响，其运行时间依赖于估计一个集合的期望影响所需要的蒙特卡洛模拟次数——$MCtimes$。当 $MCtimes$ 取值较大时，算法较为耗时；而当 $MCtimes$ 取值较小时，算法返回的解的质量难以得到保障。

基于其他方案的近似算法中，比较有代表性的是 Goldberg 等[8]提出的线性规划随机舍入算法。该方法将影响成本最小化问题视作整数规划问题，调用求解线性规划的算法求得该整数规划问题的松弛解，并对松弛解进行随机舍入，从而求得最终解。这类方法的共性是不需要目标函数满足单调性和次模性。但是，这类方法适用的模型与常见的线性阈值模型和独立级联模型有很大的区别，因此不适用于本文的情形。

3 问题描述

在相关研究中，社交网络常被建模为图 $G=(V,E)$，其中 V 是网络中节点的集合，E 是网络中边的集合。$\forall v \in V$，$c(v)$ 代表初始选择节点 v 需要付出的成本。$\forall e(u,v) \in E$，$p(u,v)$ 代表节点 u 激活节点 v 的概率。在初始时选择节点 v，网络中最终被激活的节点数量被称作 v 的影响力，记作 $I(v)$。

成本和影响的概念可从节点推广到集合。对于集合 S，其成本 $c(S)$ 为集合 S 中所有节点的成本之和，$I(S)$ 代表初始选择节点集合 S 所能激活的节点数量。

表 1 常用符号

Table 1 Frequently used notations

符号	描述
$G=(V,E)$	社交网络 G，点集 V，边集 E
n,m	图 G 中点的数量 n，边的数量 m
J	最终期望达到的影响力
$I(S)$	初始节点集合 S 的影响力
$\boldsymbol{E}(\cdot)$	随机变量的期望
$\mathcal{R}$	采样获得的反向可达集构成的集合
θ	$\mathcal{R}$ 中反向可达集的数量
$F_{\mathcal{R}}(S)$	$\mathcal{R}$ 中与 S 相交的反向可达集的比例

定义 1(影响成本最小化问题) 对于社交网络 $G=(V,E)$，给定参数 J，选择满足下式的初始节点集合 S^*：

$$S^* = \arg\min_S c(S)$$

$$\text{s.t. } \boldsymbol{E}(I(S)) \geqslant J$$

其中，约束条件要求候选集合 S 的期望影响不比 J 小，最小化的目标是 S 的成本。同时满足这两个条件的集合即为 S^*。

以图 1 中的美国大选为例，若特朗普希望在纽约州达到 55%的支持率，以获取该州的全部选票，他可在纽约州社交网络中将 J 设为 $0.55 \times n$，从而寻找一个成本最小的初始集合。

在相关研究中，常用独立级联模型来描述网络上的扩散过程。该模型假设节点 u 激活其邻接节点 v 的概率为 $p(u,v)$，任意节点都只有一次机会尝试激活其邻接节点，直到网络中不再有新的节点被激活时扩散过程结束。

下面在独立级联模型下分析问题的难度。

定理 1 独立级联模型下，影响成本最小化问题是 NP 难的。

证明：对于一个节点集可被划分为 A,B 两个集合的二部图，$|A|=m$，$|B|=n$，如果节点 $i \in A$，节点 $j \in B$，且 i,j 之间存在一条有向边，则令 $p(i,j)=1$。$\forall i \in A$，构造集合 S_i；$\forall j \in B$，构造元素 u_j。若 $p(i,j)=1$，则 $u_j \in S_i$。这样，该二部图可在多项式时间内变换为一个集合覆盖问题的实例，其基础集合为 $U=\{u_1,u_2,\cdots,u_n\}$，集合族为 $S=\{S_1,S_2,\cdots,S_m\}$。设该二部图代表一个社交网络，每个节点的代价都是 1，则判定激活全部节点的最小成本是否小于 k 等价于判定对应的集合覆盖问题是否存在小于 k 的解。因此，独立级联模

型下影响成本最小化问题的难度不低于集合覆盖问题，是NP 难的。

定理 1 说明，精确地求解影响成本最小化问题面临很大的困难。因此，只能通过设计近似算法来求解该问题，例如每次选择边际收益最大的节点。但定理 2 说明，如果通过每次选择边际收益最大的节点来设计近似算法，在服从幂律分布的网络中将会消耗大量不必要的成本。

定理 2　在服从幂律分布的网络中，每次选择边际收益最大的节点，最终返回的可行解的成本接近最优解的 $\ln(n)$倍（n 为网络中节点的数量）。

证明：在图 2 所示的服从幂律分布的网络中，设根节点激活叶子节点的概率为 1，而叶子节点无法激活其他节点。节点的激活成本被标记在图中对应的节点上。将根节点标记为节点 0，叶子节点从左到右标记为节点 1，2，…，$n-1$。利用数学归纳法：

1）对于节点 0，其成本为 $1+\varepsilon$，收益为 n，因而单位成本对应的收益为 $n/(1+\varepsilon)$；而叶子节点中节点 $n-1$ 的单位成本对应的收益最大，为 n。因此，算法第 1 轮将选择节点 $n-1$。

2）假设算法前 i 轮已经选择了节点 $n-1,n-2,\cdots,n-i$，则在算法的第 $i+1$ 轮，节点 0 的成本仍为 $1+\varepsilon$，但收益缩减为 $n-i$，因而单位成本对应的收益为 $(n-i)/(1+\varepsilon)$；而叶子节点中节点 $n-i-1$ 的单位成本对应的收益最大，为 $n-i$。因此，算法第 $i+1$ 轮将选择节点 $n-i-1$。

因此，为了在网络中达到至少为 n 的期望收益，算法将顺序地选择节点 $n-1,n-2,\cdots,0$。由调和级数的性质知，成本为 $\ln(n)$。而最优解仅选择节点 0 即可达到目标，成本为 $1+\varepsilon$。因此，算法返回的可行解的成本恰好是最优解的 $\ln(n)$倍。

图 2 所示的网络结构广泛蕴含于常见社交网络中，是一种相当常见的子结构[9]。因此，每次选择边际收益最大的节点，可行解的成本接近最优解的 $\ln(n)$倍。

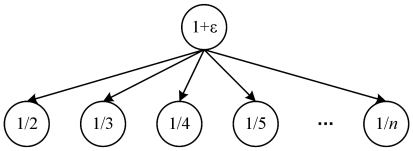

图 2　一种服从幂律分布的社交网络

Fig. 2　Social network satisfying power law distribution

为了在算法的每一轮选择边际收益最大的节点，需要估计相应节点集合的期望影响。现有研究通常利用蒙特卡洛模拟方法进行估计。定理 3 证明了这类算法的时间代价非常高。

定理 3　运用蒙特卡洛模拟方法求解影响成本最小化问题的时间复杂度为 $O(Jmn \cdot MCtimes)$（各参数的含义已在上文中进行了解释）。

证明：选择节点的总轮数由 $O(J)$所控制。为了在每轮都选择边际收益最大的节点，需要进行 $O(n)$次集合的期望影响估计，进而通过比较得到一个边际收益最大的节点。若每次估计一个集合的期望影响，则需要在网络中进行 $MCtimes$ 次宽度优先搜索，而进行一次宽度优先搜索的时间代价由 $O(m)$控制。因此，现有算法的时间复杂度为 $O(Jmn \cdot MCtimes)$。

定理 2 证明了现有算法在常见社交网络中求得的解的质量存在一定的改进空间，而定理 3 证明了现有算法的运行时间较长，难以满足即时查询的需求。因此，利用现有算法求解影响成本最小化问题在解的质量和时间效率上面临一定的瓶颈，需要提出了新的算法来突破这两大瓶颈。

对此，本文引入了基于随机采样的反向可达集技术，提出一种在解的质量和时间效率上都有着优秀表现的算法——RAISE 算法（Random sampling Approach to minimize Influence diffusion cost in Social nEtwork）。随机采样方法可以大大地降低时间代价。同时，直观地看，随机采样方法降低了图 2 所示的子结构出现的概率，有助于提高解的质量。

对于网络 $G=(V,E)$中的节点 $v\in V$，其所对应的反向可达集的构建过程为[10]：

1）对于 $\forall e(u,v)\in E$，认为该边以 $p(u,v)$的概率连通，以 $1-p(u,v)$的概率不连通。利用该判据确定网络 G 中所有边的连通性。

2）考查图 G 中所有通过已连通的边可以到达节点 v 的节点，它们构成的集合即为节点 v 所对应的反向可达集。

直观地看，若节点 u 被包含在某次采样得到的节点 v 的反向可达集中，则节点 v 在扩散过程中很可能被节点 u 所激活。Tang 等[10]证明了在社交网络的信息传播过程中，初始集合与节点 v 的反向可达集相交的概率等于节点 v 在传播过程中被激活的概率。

4　RAISE 算法

本节将详细介绍 RAISE 算法的框架，并给出该算法近似比和运行时间的理论保障。RAISE 算法主要包括以下 3 个阶段。

1）成本的度量：该阶段利用聚类和降维的方式对网络中个体的成本进行合理的度量。

2）采样：该阶段随机生成一定数量的反向可达集，并将它们置于集合 $\mathcal{R}$ 中，记采样量为 θ。

3）选择节点集合：利用解决部分覆盖问题的算法[7]，选择总成本尽可能少的节点集合，使之能够覆盖 $\mathcal{R}$ 中占比 $\lfloor J+K)/n \rfloor$的反向可达集。

RAISE 算法的具体流程如算法 1 所示。其中，第 1 行对网络中节点的成本进行了度量；第 2 行计算了为了达到理论保障所需要的采样数量；第 3 行为采样阶段，在网络 G 中通过随机采样得到 θ 个反向可达集；第 4 行为选择节点集合阶段，目标是获得一个可行解 S^*。S^* 的期望影响以不小于 q 的概率达到 J。

算法 1　RAISE(G,J,K,q)

1. G. compute_cost()

2. $\theta=4\dfrac{(G\cdot n)^2}{K^2}\ln\dfrac{1}{1-q}$

3. $\mathcal{R}$=Sampling(G,θ)

4. S^*=Selecting($\mathcal{R}$,$\lfloor$(J+K)/n$\rfloor$)

5. return S^*

4.1 节将给出成本的度量方式；4.2 节将计算为了达到理

论保障所需要的最小采样数量；4.3 节将给出选择节点集合的流程，并证明算法的近似比。

4.1 成本的度量

如前所述，说服个体的困难程度与许多因素相关。然而，若对这些因素全面加以考虑，则需要收集大量的数据，这在大多数情况下是不切实际的。因此，已有的大部分研究认为不同个体的成本是相同的。

为克服这一困难，本文提出了一种合理度量成本的框架，如图 3 所示。将个体的特征分为网络结构特征和独立于网络结构的外源特征两类，其中网络结构特征可通过对网络结构的分析计算得到。该框架的具体流程可分为以下 4 步。

1)计算个体的网络结构特征，如 PageRank 排名、度中心性、距离中心性、集聚系数等。这些特征在一定程度上反映了个体在网络中的重要性。

2)考虑获取外源特征的可行性。如果获取外源特征不具备可行性，则直接对网络结构特征加以考虑；如果可行，则将外源特征和网络结构特征进行数据融合。

3)基于以上特征，可利用聚类方法(如 K-Means 聚类算法、基于密度的聚类算法等)将网络中的个体聚为多类。

4)利用主成分分析方法对网络结构特征进行降维，主成分得分代表个体的重要程度。由于个体的成本与其在网络中的重要程度正相关，因此平均主成分得分最高的一类个体将代表高成本个体，以此类推。

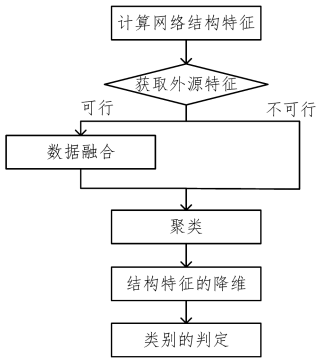

图 3 一种合理度量成本的框架

Fig. 3 Framework reasonably measuring cost of users

4.2 采样量 θ 的理论保证

为保证 RAISE 算法解的质量，采样得到的反向可达集的数量不应太小。本小节将给出采样量 θ 的理论保证。

设 $x_i(i\in[1,\theta])$ 为表征节点集合 S 是否覆盖反向可达集 R_i 的随机变量，即当 $S\cap R_i=\emptyset$ 时，$x_i=0$；$S\cap R_i\neq\emptyset$ 时，$x_i=1$。Tang 等[10]证明了节点集合期望影响的无偏估计满足下式：

$$\boldsymbol{E}(I(S))=\frac{n}{\theta}\boldsymbol{E}(\sum_{i=1}^{\theta}x_i) \tag{1}$$

因此，为了使集合 S 的期望影响不小于 J，即 $\boldsymbol{E}(I(S))\geqslant J$，只需要保证下式成立：

$$\boldsymbol{E}(\sum_{i=1}^{\theta}x_i)\geqslant J\ \frac{\theta}{n} \tag{2}$$

然而，在求解过程中，由于 x_i 满足的分布的参数未知，无法直接计算 x_i 之和的期望，因此只能对 x_i 之和的期望进行估计。定理 4 证明了在采样量充分大的情况下，x_i 之和的期望以大概率满足式(2)。

定理 4 对于任意 $\delta\in(0,1)$，$K>0$，记 $\mathcal{R}$ 中与 S 相交的集合所占的比例为 $F_{\mathcal{R}}(S)$。若：

$$nF_{\mathcal{R}}(S)\geqslant J+K \tag{3}$$

令：

$$\theta_1=-\frac{4n^2}{K^2}\ln\delta \tag{4}$$

则当 $\theta>\theta_1$ 时，$\boldsymbol{E}(I(S))$ 以 $1-\delta$ 的概率不小于 J。

证明：式(3)等价于：

$$\sum_{i=1}^{\theta}x_i-\boldsymbol{E}(\sum_{i=1}^{\theta}x_i)\geqslant K\ \frac{\theta}{n} \tag{5}$$

由于 $x_1,x_2,\cdots,x_\theta$ 为独立泊松实验，记 $Pr[x_i=1]=p_i$ $(0<p_i<1)$。记 x_i 之和为随机变量 x，其期望为 μ，对于 $\forall c>0$，下述不等式成立：

$$Pr[x-\mu>c\mu]<\mathrm{e}^{-min(\frac{c^2}{4},\frac{c}{2})\mu} \tag{6}$$

其即为切诺夫界[11]。利用该不等式，x_i 之和偏离其期望的概率满足下式：

$$Pr[\sum_{i=1}^{\theta}x_i-\boldsymbol{E}(\sum_{i=1}^{\theta}x_i)>K\ \frac{\boldsymbol{E}(\sum_{i=1}^{\theta}x_i)}{n\rho}]<\mathrm{e}^{-\frac{-\frac{K^2}{n\rho}}{4}\theta\rho} \tag{7}$$

其中，$\rho=\boldsymbol{E}(I(S))/n$。

由于 $\rho<1$，因此式(7)中的界不大于 $\mathrm{e}^{-\frac{K^2\theta}{4n^2}}$。这个界只依赖于事先选取的常数 K。因此，式(7)代表的事件未发生时，S 的期望影响不小于 J：

$$\boldsymbol{E}(\sum_{i=1}^{\theta}x_i)>\sum_{i=1}^{\theta}x_i-K\ \frac{\theta}{n}=\frac{\theta}{n}(nF_R(S)-K)\geqslant\frac{\theta}{n}J \tag{8}$$

令：

$$\delta=\mathrm{e}^{-\frac{K^2\theta}{4n^2}} \tag{9}$$

解出：

$$\theta=-4n^2\ln\delta/K^2 \tag{10}$$

因此，当 $\theta>\theta_1$ 时，$\boldsymbol{E}(I(S))$ 以 $1-\delta$ 的概率不小于 J。图 4 是采样数量关于 K/n 的图像。从图 4 中看出，RAISE 算法有一个相当明显的优点：当 $n/K=O(1)$ 时，所需的采样数量不依赖于网络的节点数量。对于包含百万量级节点数甚至规模更大的社交网络，只要 J 不是充分地接近 n，总可以选择合适的 K，使得 $n/K=O(1)$，从而保证采样数量在数百的量级。仍以美国大选为例，$J\approx0.5n$，若取 $K\approx0.2n$，则当采样量不小于 460 时，RAISE 算法有 99%的概率返回期望影响不小于 J 的节点集合。

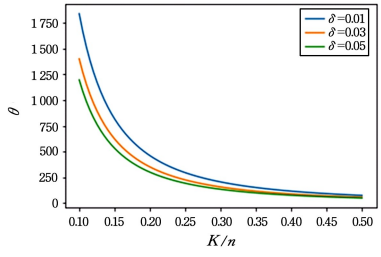

图 4 采样数量关于 K/n 的图像

Fig. 4 Images on sampling numbers versus K/n

定理 4 计算了为获得理论保障所需的采样量。定理 5 据此给出了采样阶段的时间复杂度。

定理 5　设 δ 为常数，则采样阶段的时间复杂度为 $O(mn\times \boldsymbol{E}(I(u))/K^2)$，其中 $\boldsymbol{E}(I(u))$ 为在网络中随机选取一个节点的影响力的期望。

证明：Tang 等[10]证明了：

$$\frac{n}{m}EPT=E(I(u)) \tag{11}$$

其中，EPT 代表随机采样得到一个反向可达集的时间代价的期望，即在宽度优先搜索过程中访问到的边数量的期望。因此，整个采样过程中访问到的边数量的期望为：

$$\theta\cdot EPT=\frac{\theta m}{n}\boldsymbol{E}(I(u))=-\frac{4mnln\delta}{K^2}E(I(u)) \tag{12}$$

因此，采样的时间代价为 $O(mn\times E(I(u))/K^2)$。

定理 5 中，$\boldsymbol{E}(I(u))$ 强烈依赖于网络的结构。Cha 等[12]发现在典型的社交网络中影响的传播链长一般不超过 4，而 Dinh 等[13]发现在 Facebook 等典型社交网络中消息的转发链长一般不超过 2。因此，在典型的社交网络中 $\boldsymbol{E}(I(u))$ 的数值通常不大，在 $n/K=O(1)$ 的情况下，采样阶段的时间复杂度可由 $O(n)$ 控制。若仅考虑 $\boldsymbol{E}(I(u))$ 的平凡上界，即 $\boldsymbol{E}(I(u))$ 不会超过 n，则在 $n/K=O(1)$ 的情况下，采样阶段的时间复杂度为 $O(m)$。

4.3　选择节点集合

利用采样过程中得到的反向可达集建立倒排索引，就得到每个节点可以覆盖的反向可达集。

定义 2(可覆盖集)　设 u 是图 $G=(V,E)$ 的一个节点，即 $u\in V$。称集合 $C(u)$ 为 u 的可覆盖集，对于 $\forall R_i\in R$，若 $u\in R_i$，则 $R_i\in C(u)$；若 $u\notin R_i$，则 $R_i\notin C(u)$。$C(u)$ 的成本定义为其对应节点 u 的成本，即 $C(u).cost=u.cost$。

此时，原问题转化为寻找总成本尽可能小的节点集合以覆盖给定比例的反向可达集。这是一个典型的部分覆盖问题：给定反向可达集构成的集合 $\mathcal{R}$、每个节点的可覆盖集构成的集合族 S、参数 $0<p<1$，寻找 S 的一个总成本最小的子集，使之覆盖 $\mathcal{R}$ 中元素的比例至少为 p。

为求解该部分的覆盖问题，调用 Slavik 的贪心定价法[14]，如算法 2 所示。

算法 2　Selecting ($\mathcal{R}$, p)

1. $S^*=\emptyset$
2. $r=\lceil p\times\mathcal{R}.size\rceil-|\bigcup_{u\in S^*}C(u)|$
3. if $r<0$, then return S^*
4. for u not in S^*
5. $U=\arg_u\min\frac{\mu\cdot cost}{\min(r,|c(u)|)}$
6. $S^*=S^*\cup\{v\}$
7. for u not in S^*, $C(u)=C(u)-C(v)$
8. goto 2

Slavik 证明了该贪心定价法的近似比为 $O(H(\lceil p\theta\rceil))$，其中 $H(\cdot)$ 为调和级数[14]。定理 6 利用这个结论证明了 RAISE 算法的近似比。

定理 6　设 δ 为常数，则 RAISE 算法的近似比为 $O(\ln(n/K))$。当 $n/K=O(1)$ 时，RAISE 算法的近似比为 $O(1)$。

证明：在表达式 $H(\lceil p\theta\rceil)$ 中，代入 $p=\lceil(J+K)/n\rceil$ 和 $\theta=-4n^2\ln\delta/K^2$，得到：

$$\begin{aligned}H(\lceil p\theta\rceil)&=H\left(\left\lceil\frac{J+K}{n}\frac{4n^2}{K^2}\ln\frac{1}{\delta}\right\rceil\right)\\&=H\left(\left\lceil 4\left(\frac{n}{K}+\frac{nJ}{K^2}\right)\ln\frac{1}{\delta}\right\rceil\right)\\&<H\left(\left\lceil 8\left(\frac{n}{K}\right)^2\ln\frac{1}{\delta}\right\rceil\right)\end{aligned} \tag{13}$$

利用调和级数的性质：

$$\lim_{n\to\infty}(H(n)-\ln(n))=\gamma \tag{14}$$

得到算法的近似比为：

$$\begin{aligned}O\left(H\left(\left\lceil 8\left(\frac{n}{K}\right)^2\ln\frac{1}{\delta}\right\rceil\right)\right)&=O\left(\ln\left(\left\lceil 8\left(\frac{n}{K}\right)^2\ln\frac{1}{\delta}\right\rceil\right)\right)\\&=O\left(\ln\left(\frac{n}{K}\right)\right)\end{aligned} \tag{15}$$

因此，当 $n/K=O(1)$ 时，RAISE 算法达到了 $O(1)$ 的近似比。

定理 6 说明，当 J 和 K 的量级与 n 可比拟时，算法可以达到一个常数近似比。例如，若取 $J\approx K\approx 0.2n$，$\delta=0.05$，则算法的近似比约为 5。这种情况有着深刻的现实意义与广泛的应用价值。例如，在大选中达到一定的支持率、在市场中达到一定的占有率、在网络中达到“相变”的临界点，这都要求 J 和 K 的量级与 n 可比拟。这种情况下，RAISE 算法的常数近似比即可保证花费的成本不会显著地超过实际的最小成本。

同时，对算法 2 进行分析，可以得出选择节点阶段的时间复杂度。

定理 7　选择节点阶段的时间复杂度为 $O(\theta\times\sum_{R\in\mathcal{R}}|R|)$。当 $n/K=O(1)$ 时，选择节点阶段的时间复杂度为 $O(n)$。

证明：算法 2 的第 4，5 行中，为求得获得单位收益所需成本最小的节点，需要对 $\mathcal{R}$ 中所有反向可达集涉及到的节点进行评估；再根据选择的轮数不超过 θ，即可证得选择节点阶段的时间复杂度。当 $n/K=O(1)$ 时，采样量 θ 可视作常数，而 $\mathcal{R}$ 中所有反向可达集涉及到的节点不超过网络的规模，因此时间复杂度可由 $O(n)$ 控制。

综合定理 5 和定理 7，可证得 $n/K=O(1)$ 时，RAISE 算法具备线性时间复杂度。

定理 8　当 $n/K=O(1)$ 时，RAISE 算法的时间复杂度为 $O(m)$，其中 m 为网络中边的数量。

证明：当 $n/K=O(1)$ 时，采样阶段的时间复杂度为 $O(m)$[15]，选择节点阶段的时间复杂度为 $O(n)$。因此，可将 RAISE 算法的时间复杂度写作 $O(m)$，即具备线性时间复杂度。

5　实验

在 Windows10 系统下进行实验，CPU 为 Intel Core i7-6700HQ，主频为 2.60 GHz，内存为 16 GB；所有代码采用 C/C++ 语言编写。

5.1　实验设定

1)数据集。本次实验主要使用了两个数据集，即 Net-

HEPT 和 Loc-Brightkite。

表 2 数据集的统计特征

Table 2 Statistics of real datasets

数据集	NetHEPT	Loc-Brightkite
节点数	15229	58228
边数	31376	214078
外源特征	无	有

NetHEPT 数据集是一个高能物理学家发表论文的合作网络，常用于本领域的研究中。Loc-Brightkite 是一个基于地理位置的社交网络，用户通过签到来共享他们的位置。该数据集中包含了从 2008 年至 2010 年期间的 4491143 条签到数据。以每个用户签到最为频繁的地点代表该用户的地理位置，可利用该数据集研究图 1 中的美国的政治选举。

2)对比算法。本实验使用 RAISE 算法与下列基线方法进行比较：①基于度中心性的启发式算法(degree)；②基于距离中心性的启发式算法(centrality)；③随机选择节点(random)；④J-Min-Seed Greedy 算法。

3)参数设置。实验将用户分为 3 类，即高成本用户、中等成本用户和低成本用户。高成本用户的成本在[5,10]的范围内随机选取，中等成本用户的成本在[1,5]的范围内随机选取，低成本用户的成本在[0,1]的范围内随机选取。*MCtimes* 设置为 10000 次，这是相关论文中通行的做法。网络中边 $e(u,v)$ 被赋予 $1/d_v$ 的权重，其中 d_v 是节点 v 的入度。设 J/n 的取值集合为{0.1,0.3,0.5,0.7,0.9}。

5.2 与 J-Min-Seed Greedy 算法的对比

定理 3 曾证明每轮选择边际收益最大节点的算法的时间复杂度相当高，例如 J-Min-Seed Greedy 算法[5]。在大规模网络中运行这类算法几乎没有可行性。在本部分对比实验中，采用了一个拥有 3783 个节点、24186 条边的小规模数据集——soc-sign-bitcoin-alpha，设定目标为激活网络中 10% 的节点。实验结果如表 3 所列。

表 3 soc-sign-bitcoin-alpha 数据集的对比实验结果

Table 3 Experiment results on soc-sign-bitcoin-alpha

算法	成本	耗时/s
RAISE	32	0.31
centrality	123	250.12
degree	181	229.60
random	171	141.67
J-Min-Seed	80	719522

与 RAISE 算法的对比实验表明，J-Min-Seed Greedy 算法的耗时相当长，而取得的解的质量又不优于 RAISE 算法。因此，在后文的对比实验中不再与 J-Min-Seed Greedy 算法进行对比。

5.3 解的质量

图 5 和图 6 分别给出了 NetHEPT 数据集和 Loc-Brightkite 数据集上各算法返回的可行解的成本。可以看出，RAISE 算法返回的可行解的成本明显小于各基线方法的成本。度中心性启发式算法与距离中心性启发式算法的表现较为相近，在 NetHEPT 数据集上差于随机选择节点的方法，但在 Loc-Brightkite 数据集上却优于随机选择节点的方法。这是由于 Loc-Brightkite 网络中存在大量度中心性高的节点，且这些节点之间的联系紧密；而 NetHEPT 网络中度中心性高的节点间的联系不够紧密。这体现出各启发式算法的表现在不同类型的网络中存在较大的差异。而 RAISE 算法在不同类型的网络中都有良好的表现。

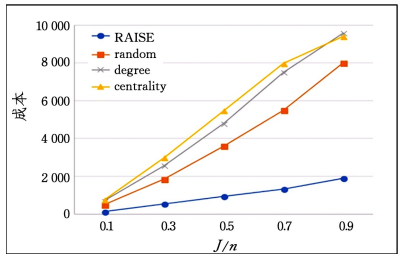

图 5 NetHEPT 数据集上的成本

Fig. 5 Cost on NetHEPT dataset

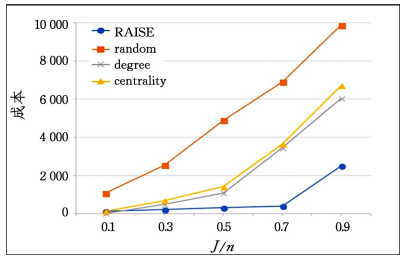

图 6 Loc-Brightkite 数据集上的成本

Fig. 6 Cost on Loc-Brightkite dataset

5.4 时间效率

图 7 和图 8 分别给出了 NetHEPT 数据集和 Loc-Brightkite 数据集上各算法的运行时间。可以看出，启发式方法利用蒙特卡洛模拟方法来估计给定节点集合的期望影响，运行时间的数量级远高于 RAISE 算法。因此，在更大规模的网络中，基于蒙特卡洛模拟的算法由于运行时间过长，将失去实用性。而 RAISE 算法的采样数量不随网络规模的变化而变化，能够满足即时查询的要求。

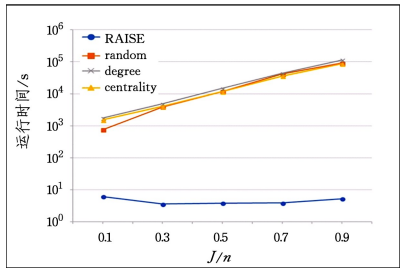

图 7 NetEPHY 数据集上的运行时间

Fig. 7 Running time on NetEPHY dataset

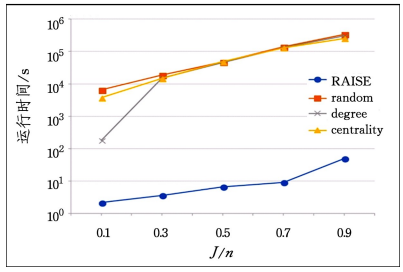

图 8 loc-Brightkite 数据集上的运行时间

Fig. 8 Running time on loc-Brightkite dataset

结束语　本文提出并形式化定义了影响成本最小化问题，也证明了它的难度；利用聚类和降维的方法给出了一种合理的成本度量方式。之后，基于随机采样方法，提出了解决该问题的 RAISE 算法；从理论上证明了在期望影响到的节点数量与网络规模可比拟的情况下，RAISE 算法具备常数近似比和线性的时间复杂度；从实践上验证了该算法在解的质量和时间效率上都优于现有方法。

参 考 文 献

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

[2] BUDAK C，AGRAWAL D，ABBADI A. Limiting the Spread of Misinformation in Social Networks[C]// Proceedings of the 20th International Conference on World Wide Web. New York：ACM，2011：665-674.

[3] LI Y，BAO Z，LI G，et al. Real Time Personalized Search on Social Networks[C]// 2015 IEEE 31st International Conference on Data Engineering (ICDE). New York：IEEE Press，2015：639-650.

[4] CHEN N. On the Approximability of Influence in Social Networks[J]. Siam Journal on Discrete Mathematics，2009，23(3)：1400-1415.

[5] LONG C，WONG R. Minimizing Seed Set for Viral Marketing[C]// IEEE International Conference on Data Mining. New York：IEEE Press，2011：427-436.

[6] ZHANG P，CHEN W，SUN X，et al. Minimizing Seed Set Selection with Probabilistic Coverage Guarantee in a Social Network[C]// Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York：ACM，2014：1306-1315.

[7] KEMPE D，KLEINBERG J，TARDOS E. Maximizing the Spread of Influence through a Social Network[C]// Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM，2003：137-146.

[8] GOLDBERG S，LIU Z. The Diffusion of Networking Technologies[C]// Twenty-fourth Acm-siam Symposium on Discrete Algorithms. New York：SIAM，2013：1577-1594.

[9] BARABASI A，ALBERT R. Emergence of Scaling in Random Networks[J]. Science，1999，286(5439)：509-512.

[10] TANG Y，XIAO X，SHI Y. Influence Maximization：Near-optimal Time Complexity Meets Practical Efficiency[C]// Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York：ACM，2014：75-86.

[11] MOTWANI R，RAGHAVAN P. Randomized Algorithms[M]. Cambridge：Cambridge University Press，1995：67-73.

[12] CHA M，MISLOVE A，GUMMADI K. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network[C]// Proceedings of the 18th International Conference on World Wide Web. New York：ACM，2009：721-730.

[13] DINH T，NGUYEN D，THAI M. Cheap，Easy，and Massively Effective Viral Marketing in Social Networks：Truth or Fiction?[C]// Proceedings of the 23rd ACM conference on Hypertext and social media. New York：ACM，2012：165-174.

[14] SLAVIK P. Improved Performance of the Greedy Algorithm for Partial Cover[J]. Information Processing Letters，1997，64(5)：251-254.

[15] TANG Y，SHI Y，XIAO X. Influence Maximization in Near-linear Time：A Martingale Approach[C]// Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. New York：ACM，2015：1539-1554.