文章编号:1007-5321(2019)03-0083-08 DOI:10.13190/j.jbupt.2018-204

基于通联行为的信息传播模式挖掘方法

项英倬, 魏 强, 游 凌
(盲信号处理国家重点实验室, 成都 610041)

摘要: 针对通信内容未知且无关通联占比高情况下信息传播模式的挖掘问题,提出了一个生成模型,对通联行为发生的时间建模,预测网络中用户通信内容的相关性,进而获取网络中信息的传播模式. 证明了求解所提模型的复杂度为 NP-hard,并提出用 NetMine 算法来估计模型的一个近似最优解. 实验结果表明,所提 NetMine 算法能够高效地挖掘网络中信息的传播模式,并优于已知的其他方法.

关 键 词: 信息传播; 数据挖掘; 信息流; 子模函数
中图分类号: TP311 **文献标志码**: A

An Information Diffusion Pattern Mining Method Based on Communication Actions

XIANG Ying-zhuo, WEI Qiang, YOU Ling
(National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu 610041, China)

Abstract: To deal with the challenges of information diffusion pattern mining problem which the communication content is unknown and innocent data occupies a very high ratio of the observed data, the article proposes a probability model predicting the relativity of the communications between users, which infers the information diffusion. In addition, it proves the inferring problem NP-hard, and proposes NetMine algorithm to get a near optimal solution. Experiments show that the proposed NetMine algorithm outperforms other state-of-art algorithms.

Key words: information diffusion; data mining; information flow; submodular function

信息的传播、影响力的分析、木马和病毒的传播等是社交网络和信息网络中重要的研究领域[1-3],挖掘网络中信息传播模式将有助于这些问题的分析. 网络中信息的传播模式指目标信息(木马、病毒、某个观点等)在网络中传播的模式,该模式难以通过观察得到. 挖掘网络中信息的传播模式需要解决 2 个挑战:① 网络中传播的信息内容是未知的、无法获取的;② 网络中背景通信流量占比远高于目标信息流量. 一般的场景下,能够利用的数据只有通联双方和通信发生的时间. 比如,在挖掘僵尸网络中,比较容易获取节点间通信行为及通信时间,但节点间通信的负载难以进行分析. 目前,关于信息传播的研究方法大都基于通信内容的方法[4-5]、通信内容与通信属性结合的方法[6-8],或者是基于复杂网络的一些方法[9-10]. 而这些方法在通信内容未知且通信属性只有时间时,难以发挥很好的效果.

笔者所研究的场景比较特殊,在仅知网络中用户通联双方及通联时间的情况下,推断用户间传递

收稿日期: 2018-11-01
基金项目: 国家自然科学基金项目(61174124)
作者简介: 项英倬(1990—), 男, 博士生.
通信作者: 魏 强(1987—), 男, 助理研究员, E-mail:weiqianglg@163.com.

内容的相关性,进而分析用户间信息的传播模式. 通常情况下,容易观测到某用户与其他用户发通信的时间序列,而通信内容未知,那么,如何通过这些通信数据分析网络中的信息是如何传播的呢? 网络中什么样的信息传递模式能够来解释这些观测的通联数据呢? 为了解决这些问题,采用生成模型对用户间通联数据的相关性和信息传播模式进行建模,并试图找到一个隐含的网络子图,使得通联数据的似然函数能够最大化. 网络中的指令、僵尸网络中的指控数据等信息的传播存在一个固定的模式,如果网络中通联数据能够构成一个有向多边网络(multi-edge digraph),信息的传播模式将是该网络的一个有向子图.

1 信息传递模式挖掘方法

1.1 问题定义

一般来说,网络中用户间信息的传播遵循一定的模式,而这个模式通常可以表示用户间信息传递的关系. 下面给出网络中用户间信息传播模式的一个定义.

定义1 网络中用户间信息传播模式指在通信网络 $G_o(V, E)$ 中代表用户间信息传播的一种固有模式. 本节中用网络 $G(V, E)$ 代表用户间信息传播关系的网络,那么 $G(V, E)$ 是一个有向图,且是 $G_o(V, E)$ 的一个子图,代表用户间的信息更加倾向于沿着网络 $G(V, E)$ 中的边传播. 网络 $G(V, E)$ 中的每条边称为信息传递边,而网络 $G_o(V, E)$ 的非信息传递边称为普通边.

定义2 用户间信息传递模式挖掘问题指给定节点的集合 V,以及节点之间的通联数据 $A = \{(v_i, v_j, t_k) \mid v_i, v_j \in V, i \neq j\}$,如何推测节点间传递信息的相关性,并得到 $G(V, E)$.

容易得到通信网络 $G_o(V, E)$ 的拓扑,只需要将每条通联数据当作一条边,并将所有的边组合起来;而用户间信息的传递模式 $G(V, E)$ 是不容易得到的. 由于通信内容是未知的,为了得到网络中信息的传播模式,首先需要根据节点的通信行为,推测通信内容的相关性,然后分析信息的传播模式. 文献[11-13]中指出,信息转发的时间间隔 Δt 满足指数分布或幂律分布,即 $p(\Delta t) \sim e^{-\alpha \Delta t}$ 或者 $p(\Delta t) \sim (\Delta t)^{-\alpha}$. 如果用户收到某次通信后在短时间内又与其他用户通信,那么可以推测用户间传递的信息是相关的,这里使用 $f(t_{i,j} \mid t_{k,i})$ 来衡量传递信息的相关性.

1.2 信息传播模型

为了方便后续数据处理,首先需要将通联数据 A 进行整理. 这里定义用户的行为记录.

定义3 用户的行为记录指用户接收信息与发送信息的行为所构成的记录. 那么,$R_{\text{record}} = \{R_i \mid i \in V\}$,其中

$$R_i = \left\{ \begin{matrix} \cdots r_x r_y \cdots s_p \cdots \mid r_x := \text{recievetime of node } x \rightarrow i, \\ s_p := \text{sendtime of node } i \rightarrow p \end{matrix} \right\} \tag{1}$$

定义一个最大观测时间窗口 T,并将每个用户的行为记录切成许多时间切片 ts. 由于需要挖掘用户传递信息的行为,所以仅从用户收到其他用户发送给其信息的时间开始作为时间切片的起点,并只考虑用户的发送行为. 笔者认为,用户在时间窗口外的发送行为与窗口内收到信息的时间间隔过长而不会存在信息传递的可能. 图1为用户 i 的行为记录及时间切片的示意图. 在时间切片 r_x 与 $r_x + T$ 之间只有2个发送行为,因此该时间切片可以表示为 $\text{ts}_{r_x,i} = \{r_x, s_u, s_v\}$,每个时间切片由2个下标来确定,第1个下标代表该时间切片开始的时间,第2个下标代表用户 i 的时间切片,并且每个时间切片的第1个数据为该用户收到信息的用户及时间,后续的数据全部为该用户的发送行为及时间. 切割所有用户的行为记录,将通联数据 A 转化为了时间切片的集合 TS,容易证明,通联数据 A 与时间切片 TS 是可以互相转化的,时间切片只是对数据变换了一个表现形式.

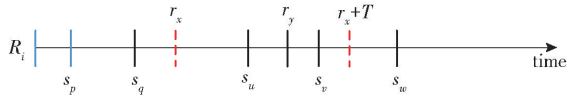


图1 节点行为记录及时间切片示意图

在图1中,由于并不知道用户 i 收到来自用户 x 的消息后,是否将消息转发给了用户 u 或者 v,因此需要根据不同情况进行讨论. 如果由用户 x 指向用户 u 之间的边属于 $G(V, E)$,那么用户有很大的概率是进行信息传播的行为,使用 $f(t_{i,u} \mid t_{x,i})$ 来表示衡量这种概率的大小;如果由用户 x 指向用户 v 之间的边不属于 $G(V, E)$,那么可以使用一个非常小的数值 ε 来表示这种情况发生的概率. 由于用户转发消息时可能会发送给 $G(V, E)$ 中一个或多个子节点,那么用 β 表示用户未发送子节点的概率大小.

这样可以得到一个时间切片的似然函数：

$$f(\mathrm{ts}_{r_x,i}\mid G,G_o)=\prod_{k\in Vout(i)}\varepsilon^m(1-\varepsilon)^n\beta^q f(t_{i,s_k}\mid t_{r_x,i})=\prod_{k\in V_{\mathrm{out}}(i)}f'(t_{i,s_k}\mid t_{r_x,i},\varepsilon,\beta)\tag{2}$$

其中：t_{i,s_k}为用户 i 发送信息给用户 k 的时间，n 为时间切片中用户 i 发送信息给子节点的通信数量，m 则为用户 i 发送信息给非子节点的通信数量，q 为用户未发送给其子节点的数量，$m+n$ 为用户 i 在 $G_o(V,E)$ 中的出度与 $G(V,E)$ 中出度的差，V_{out} 代表节点的出射边集合.

图 2 给出了式(2)的一个说明. 图中虚线表示 $G_o(V,E)$ 中的非信息传递边，实线表示 $G(V,E)$ 中的信息传递边. 一般来说，用户间信息传递模式 $G(V,E)$ 是未知的，可以发现，第 1 个潜在传递模式更加容易产生右边的时间切片，因此由式(2)表征的似然函数就要大一些.

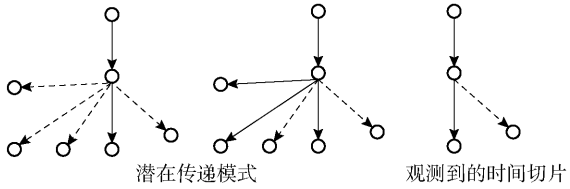


图 2　时间切片与潜在传播模式

考虑所有的时间切片集合 TS，可以得到全部观测数据的似然函数为

$$f(\mathrm{TS}\mid G,G_o)=\prod_{i\in V,r_x\in R_i}f(\mathrm{ts}_{r_x,i}\mid G,G_o)\tag{3}$$

这样，将通信网络中用户间信息传递模式挖掘问题转化为了如何找到 G 使得式(3)最大化. 下面对式(3)的一些性质进行分析. 首先，式(3)是非负的，因为式(2)是非负的；其次，式(3)是单调的，也就是说，对于 $G(V,E)$ 及 $G'(V,E')$，$E\subset E'$，$f(\mathrm{TS}\mid G,G_o)<f(\mathrm{TS}\mid G',G_o)$，因为根据式(2)，将一条普通边转变为信息传递边后，似然概率中的一个 ε 变为更大的概率值. 那么，向 $G(V,E)$ 中增加边后并不会降低解的性能，$G(V,E)=G_o(V,E)$ 应该是最佳的解，然而希望挖掘出的信息传播模式 $G(V,E)$ 要稀疏一些，这样更能够表征原网络 $G_o(V,E)$ 中的一些特性，尤其是得到信息传播的一些骨干路径图. 因此，将网络 $G(V,E)$ 中边的数据限制为 k. 重新改写通信网络中用户间信息传递模式挖掘问题为

$$G'=\arg\max_{|G|\leqslant k}f(\mathrm{TS}\mid G,G_o)\tag{4}$$

一般来说，采用暴力搜索的方法求解式(4)需要指数的时间，下面证明求解式(4)是 NP-hard.

定理 1　由式(4)定义的通信网络中用户间信息传递模式挖掘问题是 NP-hard.

证明　见附录 A.

定理 1 证明了通信网络中用户间信息传递模式挖掘问题为 NP-hard，因此求解该问题难以在多项式时间内找到求解算法. 通常，求解这类问题一般采用启发式的搜索算法或者是求解一个近似最优解.

1.3　NetMine 算法

为了求解式(4)，先对其两边取对数，将似然函数 $f(\mathrm{TS}\mid G,G_o)$ 转化为 log 似然函数 $\log f(\mathrm{TS}\mid G,G_o)$. 更进一步，对等式右边做等价转化，改为优化增量最大化. 首先假设一个空网络 K，其边全部是普通边.

$$G'=\arg\max_G(\log f(\mathrm{TS}\mid G,G_o)-\log f(\mathrm{TS}\mid K,G_o))\tag{5}$$

将式(2)和式(3)代入式(5)得到

$$G'=\arg\max\sum_{i\in V,r_k\in R_i}\sum_{j\in V_{\mathrm{out}}(i)}\log w(i,j)\tag{6}$$

其中：$w(i,j)=\varepsilon^{-1}f'(t_{i,s_j}\mid t_{r_x,i},\varepsilon,\beta)$. 为了方便，定义

$$F(\mathrm{TS}\mid G)=\sum_{i\in V,r_k\in R_i}\sum_{j\in V_{\mathrm{out}}(i)}\log w(i,j)\tag{7}$$

下面研究式(7)的一些性质.

定理 2　假设 V 为节点的集合，TS 为时间切片的集合，那么 $F(\mathrm{TS}\mid G)$ 是子模函数(submodular function) F：$2^w\to\mathbb{R}$，其中，$W\subseteq V\times V$ 为有向边的集合.

证明　见附录 B

一般来说，子模函数的最大化问题都是 NP-hard[14]，但是笔者通过贪婪算法求得了一个近似最优解，从一个空图 K 开始，每次循环向图里添加一个可以使式(7)的增量最大的边，直到 K 里有 k 条边为止. 根据上述思想，下面给出了 NetMine 算法的伪代码.

算法 1　NetMine 算法

Require：TS，k，G_o

$G\leftarrow K$

for all ts ∈ TS

　　$S_{\mathrm{ts}}\leftarrow$all possible subordinates

While $|G|<k$

　　For all$(i,j)\in G_o\backslash G$

　　$\delta_{i,j}=0$，$M_{i,j}\leftarrow\varnothing$

For all ts:$(i,j) \in$ ts

If $w(i,j) \geqslant w(\text{Parent}_{\text{ts}}(j),j)$ then

$\delta_{j,i} = \delta_{j,i} + w(i,j) - w(\text{Parent}_{\text{ts}}(j),j)$

$M_{i,j} \leftarrow M_{i,j} \cup \{\text{ts}\}$

$(i^*,j^*) \leftarrow \arg\max_{(i,j)} \delta_{i,j}$

$G \leftarrow G \cup \{(i^*,j^*)\}$

For all ts $\in M_{i^*,j^*}$

$\text{Parent}_{\text{ts}}(j^*) \leftarrow i^*$

Return G

在代码实现中,还采用了局部更新和延迟计算[15]的方法来加快迭代的速度. 因为通常每个时间切片并不是特别大,一般仅涉及网络中的几个点,如果一次循环中 NetMine 算法选择的边不涉及该时间切片中任何一个节点时,那么该时间切片对下一个循环的增量贡献为零,因此在计算下一次的循环过程中,仅需要计算有影响的时间切片即可. 这样,可以大幅度减少每次循环中的计算量. 而延迟更新的思想则是:在每次循环过程中,优先从选择较大增量的边来进行比较. 假设 $G_1, G_2, \cdots, G_k$ 是贪心算法每次循环后依次得到的网络,令 $\Delta_e(G_i) = F(\text{TS} \mid G_i \cup \{e\}) - F(\text{TS} \mid G_i)$ 代表第 i 次循环中将边 e 添加到网络中后得到的增量. 由于 $F(\cdot \mid G)$ 的子模性,那么当 $i \leqslant j$,有 $\Delta_e(G_i) \geqslant \Delta_e(G_j)$. 由此,在每次循环中,边 e 的增量只能是单调递减的,这意味着在一次循环中增量较小的边不可能在下一次的循环中使得增量突然变大. 这样,每次循环都维护一个增量的序列,每次循环时,算法可以优先从最大增量的边开始,由于在新的循环中,该增量可能会降低,那么只要重新计算增量即可,并将改变重新插入到增量序列中;如果增量不变,就选取这条边作为本次循环的最优边.

NetMine 算法也比较容易进行并行化处理,从而提升计算速度及对大规模网络的适应性. 比如在初始化计算每个时间切片的似然函数及更新边的增量时,容易进行大规模的并行处理,这又可以增快算法的速度.

2 实验分析

2.1 模拟数据仿真

使用模拟数据对 NetMine 算法的性能进行分析,并与基本方法和静态图算法进行对比验证. 为了生成模拟数据,首先确定一个节点的集合 V,然后使用 Kronecker 图[16-17]生成真实的有向网络结构,用户之间的信息将在网络的有向边上传播,该网络结构通常代表用户间信息传递的传递模式. 考虑随机图[18](Kronecker 参数矩阵为[0.5, 0.5; 0.5, 0.5])、层次社区结构[19](Kronecker 参数矩阵为[0.962, 0.107; 0.107, 0.962])及随机幂律随机树[20]3 种不同的网络结构. 实验中,随机选择网络中一个节点作为信息传递的起始节点,并为每条边设置一个转发概率来控制信息传播的广度,然后设置一个转发时间间隔参数来控制用户转发信息速度的快慢. 这样,可以模拟生成一系列用户传递信息的数据,且网络中每条边尽可能参与一次信息的传播. 然后,生成大量的随机通联数据来模拟网络中用户的背景通信数据,为使模拟数据更加符合真实数据中低信噪比的情况,用户间信息传递数据与背景通信数据比值一般小于 0.01,为了方便,采用信噪比来表示这一指标.

例如在实验中,模拟生成一个具有 64 个节点及 75 条有向边的层次型社交网络,然后模拟用户的信息传递行为,生成 180 条信息传递记录及 1.2 万条背景通信数据,使得层次社交网络中全部的边均至少参与了一次信息的传递.

比较 2 个算法:来自文献[9-10]的静态图算法和基本算法. 静态图算法的基本思想是:选取通联数据中出现次数、通联性等得到的 k 个边作为挖掘结果;基础方法的思想是:采用 $p(u,v) = \sum_{\text{ts} \in \text{TS}} P_{\text{ts}}(u,v)$ 作为边的权重,即信息由边 u 传给边 v 的似然程度,并选取权重最大的 k 个边作为挖掘结果.

为了评估算法的性能,通过对比挖掘出的信息传递模式图 G 与模拟生成的网络 G' 进行对比来确定结果的好坏. 定义查全率为挖掘出的网络中的边在真实网络 G' 中的比例;准确率为挖掘出的边正确的比例. 希望的结果是在高准确率的基础上,尽可能地提高查全率,也就是说,如果挖掘出少量边,希望这些边尽可能都在真实的网络 G' 中,而随着挖掘出的边增多,准确率会逐渐下降,而查全率单调升高.

图 3 给出了 3 种算法的仿真实验结果. 从结果可以看出,信噪比为 0.03 时的算法性能要好于信噪比为 0.004 时的算法性能. 这说明信噪比对于算法的性能影响还是比较明显的. 在 3 种算法中,无论是不同的网络结构类型还是不同的信噪比,静态图

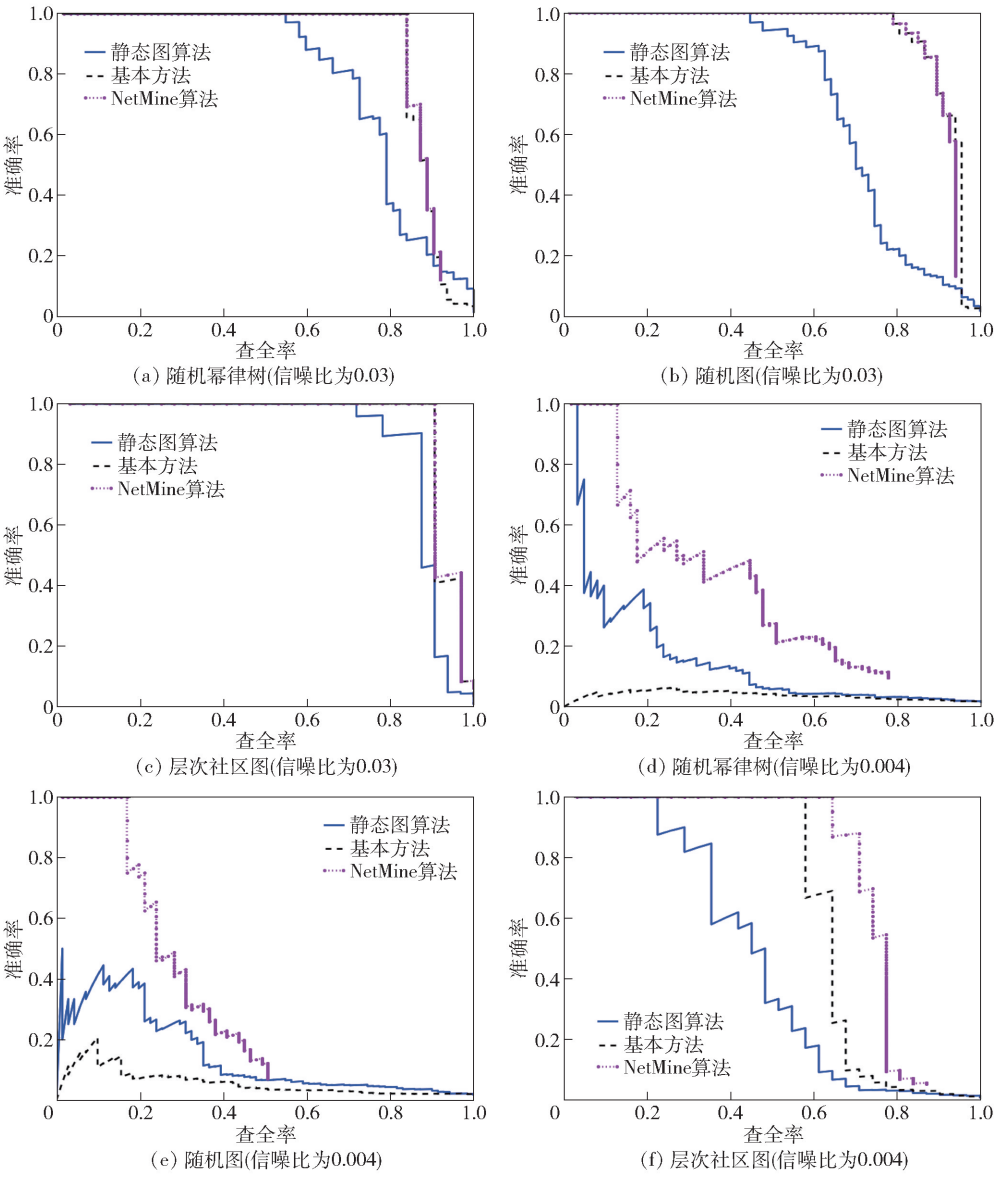


图 3　算法性能对比

算法性能都是最差的，而且当信噪比降低到 0.004 之后，该算法在随机图网络中已经失效，无法挖掘出可信的结果. 这说明了对于这种具有时序信息数据的挖掘，传统的基于复杂网络指标的挖掘算法难以挖掘出用户的信息传递模式. 通联频率高的 2 个节点之间的边未必是对信息传递影响最大的边，这对于常识里面认为通联频率高、数据量大的线路最重要有一定出入. 这些通联频率高的边可能在某种程度上说明这条线路的两个端点之间有大量的数据要传输，然而对于网络中用户间信息的传递来说，这条边可能并不是那么重要. 这说明了不同的网络结构中，由于结构的不同，有的边可能需要承载大量的数据，然而这些数据以背景通联数据为主；而用户间信息的传递可能未必会通过该边进行传播，用户间信息的传递模式网络与观测到的通联数据网络存在很大差异.

提出的 NetMine 算法在信噪比为 0.03 时与基础方法性能差别不大，都能够达到不错的性能. 这两种算法对于不同的网络结构均能够在查全率为 0.9 左右时准确率达到 0.95 以上，这说明了算法的可靠性及稳定性. 而当信噪比降到 0.004 时，基础方法已经基本失效，只有在层次社交网络结构中还具有一定性能；而 NetMine 算法在这 3 种网络结构中均可以取得最好的效果.

一般来说，如果能够获取更多的用户间信息传递数据，就能够显著提高算法性能. 而本实验中，仅

模拟生成了180条用户间信息传递数据,这能够说明NetMine算法在有效数据量不大的情况下,仍可以很好地挖掘出用户间信息传递的模式.

从不同的网络结构看,算法对于层次社交网络结构的挖掘效果最好,而随机图和随机树这2个随机网络的挖掘性能要稍微差一些. 一个可能的原因在于:随机网络中,模拟用户传递信息的通联数据与随机数据过于类似,从而导致算法难以获取有用的信息. 尽管NetMine算法性能在不同的网络结构下有一定区别,但仍然相对比较稳定,这从稳定性的角度说明了NetMine算法的可靠性.

图4给出了算法性能与数据信噪比在不同网络结构下的关系. 实验中,每个网络具有512个节点以及755条边,每次仿真中仅生成了300条用户间信息传递的数据. 从图中可以看出,固定内容相关的通联数据量之后,NetMine算法和静态图算法性能都随着信噪比的提升而提高,这在3种不同的网络结构下均表现出相同的趋势. 更进一步,从图中可以看出,NetMine算法的性能无论信噪比为何值,均优于静态图算法;NetMine算法性能随着信噪比的提升,在信噪比为0.4之后达到平稳的状态,BreakPoint大约在0.94左右,而且算法性能在3种网络结构下都非常稳定,这说明了算法挖掘结果的准确率和查全率都非常高,而静态图算法的性能上限在0.7左右,而且不同的网络结构下,其性能起伏比较大.

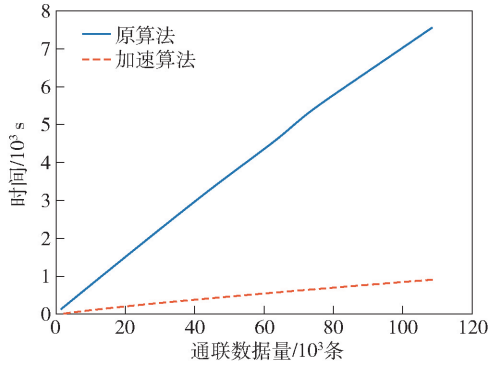
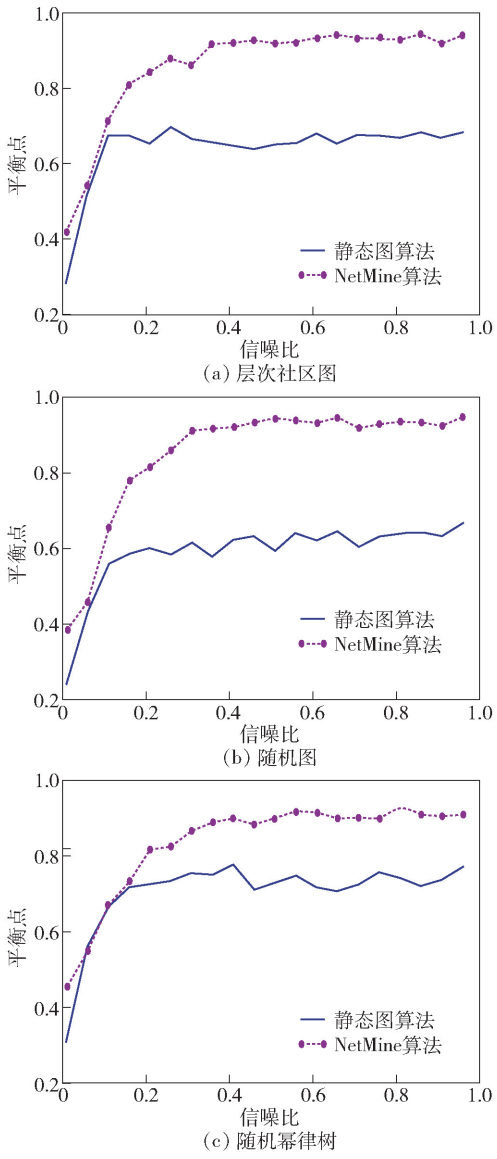


图4 算法性能与信噪比

提出了延迟计算和局部更新的技术来对算法进行加速,图5给出了算法的加速性能.

实验中采用了层次社区结构这一网络结构,图中横坐标为观测到的总数据量,纵坐标为算法运行的时长. 从图中可以看出,NetMine算法中采用的延迟计算和局部更新算法能够大幅度加快搜索最优解



(a) 层次社区图

(b) 随机图

(c) 随机幂律树

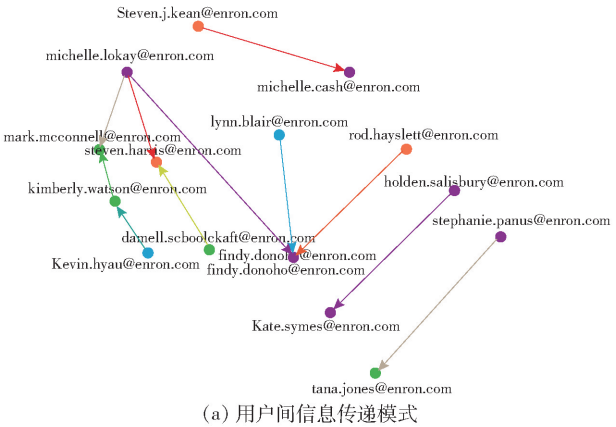
图5 算法加速性能分析

的速度,速度提升了近7倍,而且算法耗时随通联数据量的增长而增加缓慢. 相比于贪心算法,NetMine算法的耗时曲线斜率要平缓得多,处理5万条通联数据仅需要500 s左右的时间. 从图5中还可以看出,算法耗时随着通联数据量的大小几乎是线性增长,相比于多项式级别的增长来说,这将非常有利于处理大规模的通联数据.

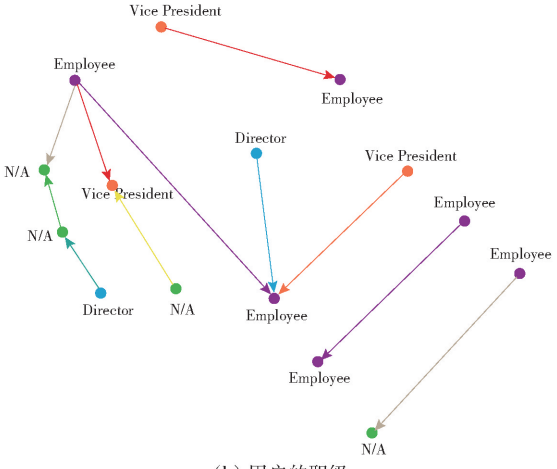
2.2 Enron邮件集数据实验

Enron邮件集[21]是Enron公司几千名员工办公邮箱中的邮件数据集合,最初由联邦能源局公开,由卡内基梅陇大学的William Cohen收集并用于科学研究[22]. 采用了其中一个含有151名标注了员工岗位职级的版本,由于仅需要邮件通信的双方和时

间，所以舍弃了邮件的内容，仅提取了邮件的发送者、接收者及邮件发送时间，将这些数据存入 MySQL 数据库中. 将有效的邮件数据输入到 NetMine 算法中，设置算法的参数 $k=11$，也就是说，仅挖掘由 11 条边构成的信息传递网络，结果如图 6 所示.



(a) 用户间信息传递模式



(b) 用户的职级

图 6　Enron 邮件集中用户信息传递模式与职级

从图 6 中可以看出，Enron 公司员工间信息的传递模式大部分是由高级职位的员工传向低职位的员工，这可以看作是一种指令类信息的传播过程，由高层向低级别员工传递信息. 但是，也不总是这样，比如 Steven Harris 这名高管，他属于信息的汇聚方，信息由下属流向了他. 研究 Steven Harris 的邮件内容和标题发现，这名高管几乎所有的邮件都是在回复他人的邮件，很少主动发送邮件，也就是说，这名高管的工作流程通常是员工将信息汇总到他这里，然后他再进行相应处理，处理后再转发出去. 因此，这是一个信息汇聚的传播模式.

从结果中看，网络中用户间信息传递的模式中每条信息传递路径的长度并不是特别长，电子信息技术的发展拉近了人与人之间的距离，从高管到员工需要的转发次数并不是太多.

3　结束语

研究了通信网络中用户间信息传递模式的挖掘问题，并提出了 NetMine 算法对用户间信息传递模式进行挖掘. 通过模拟的仿真实验发现，NetMine 算法能够较好地挖掘出相应的信息传递模式，且仅利用了通信发生的时间这一属性，就取得了出乎意料的结果. 通过与基于静态网络的挖掘算法对比发现，网络中流量大的边在信息的传播中所起的作用未必会很大，这对于研究网络中信息的传播、市场营销等具有很强的指导意义. 将算法应用于 Enron 邮件集合中，分析了 Enron 公司中员工间信息的传递模式.

进一步的研究内容将在运行速度方面对算法进行优化，使其能够更快地处理海量数据.

参考文献：

[1] Katz E, Lazarsfeld P F, Roper E. Personal influence: the part played by people in the flow of mass communications[J]. The Canadian Journal of Economics and Political Science, 1957, 23(4): 572-574.

[2] Rogers E M, Singhal A, Quinlan M M. Diffusion of innovations[C]//An Integrated Approach to Communication Theory and Research. Routledge: [s. n.], 2014: 432-448.

[3] Watts D J, Dodds P S. Influentials, networks, and public opinion formation[J]. Journal of Consumer Research, 2007, 34(4): 441-458.

[4] McCallum A, Wang Xueri, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on enron and academic email[J]. Journal of Artificial Intelligence Research, 2007, 30(1): 249-272.

[5] Zhang Yang, Wu Yao, Yang Qing. Community discovery in twitter based on user interests[J]. Journal of Computational Information Systems, 2012, 8(3): 991-1000.

[6] Fei Hongliang, Jiang Ruoyi, Yang Yuhao, et al. Content based social behavior prediction: a multi-task learning approach[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM, 2011: 995-1000.

[7] Lin C, Mei Q, Jiang Y, et al. Inferring the diffusion and evolution of topics in social communities[J]. Social Network Mining and Analysis, 2011(3): d5.

[8] Zhu Jiang, Xiong Fei, Piao Dongzhen, et al. Statistically

modeling the effectiveness of disaster information in social media[C]// Global Humanitarian Technology Conference (GHTC). New York: IEEE Press, 2011: 431-436.

[9] Altenburger K M, Ugander J. Monophily in social networks introduces similarity among friends-of-friends[J]. Nature Human Behaviour, 2018, 2(4): 284-290.

[10] Yang Ming, Hsu W H, Kallumadi S T. Predictive analytics of social networks: a survey of tasks and techniques[J]. Social Media Marketing: Breakthroughs in Research and Practice. IGI Global, 2018: 823-862.

[11] Leskovec J, McGlohon M, Faloutsos C, et al. Patterns of cascading behavior in large blog graphs[C]// Proceedings of the 2007 SIAM International Conference on Data Mining. 2007: 551-556.

[12] Barabasi A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039): 207-211.

[13] Malmgren R D, Stouffer D B, Motter A E, et al. A poissonian explanation for heavy tails in e-mail communication[J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(47): 18153-18158.

[14] Khuller S, Moss A, Naor J S. The budgeted maximum coverage problem[J]. Information Processing Letters, 1999, 70(1): 39-45.

[15] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007: 420-429.

[16] Goyal A, Bonchi F, Lakshmanan L V S. Learning influence probabilities in social networks[C]// Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 241-250.

[17] Leskovec J, Faloutsos C. Scalable modeling of real graphs using Kronecker multiplication[C]// Proceedings of the 24th International Conference on Machine Learning. New York: ACM, 2007: 497-504.

[18] Erdös P, Rényi A. On the evolution of random graphs[J]. Publ Math Inst Hung Acad Sci, 1960(5): 17-61.

[19] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98-101.

[20] Grover A, Zweig A, Ermon S. Graphite: iterative generative modeling of graphs[J]. arXiv preprint arXiv: 1803. 10459, 2018.

[21] Shetty J, Adibi J. The Enron email dataset database schema and brief statistical report[J]. Information Sciences Institute Technical Report. [S.l.]: University of Southern California, 2004, 4(1): 120-128.

[22] Cohen W W. Enron email data set[EB/OL]. 2013. http://www.cs.cmu.edu/~enron/.

附录 A 定理 1 证明.

将通信网络中用户间信息传递模式挖掘问题规约到 MAX-k-COVER 问题[14]中来. 在 MAX-k-COVER 问题中, 给定一个有限集合 $W(|W|=n)$和一些子集 $S_1, \cdots, S_m \subset W$ 的集合, 目标函数为

$$F_{MC}(A) = |\cup_{i\in A} S_i| \tag{A1}$$

式(A1)表示由 A 索引的子集覆盖 W 中元素的个数. MAX-k-COVER 的目标是选取 k 个子集使得 F_{MC}最大化. 假设一个势为 n 的时间切片集合 TS 及求解的目标网络 G, 使得$\max\limits_{|G|\leqslant k} F_{\mathrm{TS}}(G) = \max\limits_{|A|\leqslant k} F_{MC}(A)$. 网络 G 的节点为 $V=\{1, \cdots, m\}\cup\{r\}$, 可以做一个双射 $(i, r)\leftrightarrow S_i$, 那么 $f(\mathrm{TS}|G, G_o)$将会增加(如果将一个普通边转化为信息传递模式的边), 因此可以做如下映射: $f(\mathrm{TS}|G\cup(i, r))\leftrightarrow S_i$. 如果在 G 中增加一条边 $(i, r)\in G$, 则选择 S_i 进入 A, 那么一个含有 k 条边的网络 G 等价于 MAX-k-COVER 问题中的一个解 A. 因此, MAX-k-COVER 的每个解 A 都可以等效为由式(4)定义的问题的一个解 G.

附录 B 定理 2 证明.

首先考虑单个时间切片 ts, 网络 $G\subset G'$, 以及一条边 $e=(r, s)\notin G'$. 假设 $w_{i,j}$为网络 G 中边(i, j)的权重,$w'_{i,j}$为 G'中边(i, j)的权重. 因为 $G\subset G'$, 显然对于所有的边有 $w'_{i,j}\geqslant w_{i,j}\geqslant 0$. 如果$(i, j)$为网络 G 和 G'共同的边, 那么 $w_{i,j}=w'_{i,j}$. 设 $M_{A,e}=\sum\limits_{i\in A\setminus\{r\}} w(i, s)$, 容易得到 $M_{G,e}\leqslant M_{G',e}$, 因此

$$F(\mathrm{ts}|G\cup\{e\}) - F(\mathrm{ts}|G) = \log\left(\frac{M_{G,e}+w(r, s)}{M_{G,e}}\right)\geqslant\log\left(\frac{M_{G',e}+w(r, s)}{M_{G',e}}\right) = F(\mathrm{ts}|G'\cup\{e\}) - F(\mathrm{ts}|G') \tag{B1}$$

由于子模函数的非负线性组合仍然是子模函数, 所以 $F(\mathrm{TS}|G)=\sum F(\mathrm{ts}|G)$也是子模函数.