基于随机游走的大规模图中节点对采样算法*

吴春琼[1]，叶东毅[2]
（1. 福州大学阳光学院，福州 350015；2. 福州大学 数学与计算机科学学院，福州 350116）

摘　要：社会网络中的节点对采样可用于大规模社会网络的好友预测和用户兴趣识别。当整个网络的拓扑结构不完全或者随机选择用户的代价很高时，传统的均匀顶点采样方法的性能迅速下降。为此，提出了一种基于随机游走的大规模图中节点对采样算法。首先对社会网络的节点对采样进行了系统分析，对不同跳数下的节点对进行了定义；然后将社会网络转换成等价的网络图。新图中的顶点是原图中的边，新图中边的两个顶点是原图中含有相同顶点的两条边。最后，在新图上应用随机游走模型对节点对进行采样。实验结果表明，提出的方法统计误差小、执行效率高，性能明显优于均匀节点采样的相关算法。

关键词：图；随机游走；均匀顶点采样；社会网络
中图分类号：TP301.5　　**文献标志码**：A　　**文章编号**：1001-3695(2015)04-1052-04
doi:10.3969/j.issn.1001-3695.2015.04.022

Random walk based node pair sampling algorithm in large graph

WU Chun-qiong[1], YE Dong-yi[2]
(1. *Sunshine College of Fuzhou University*, *Fuzhou* 350015, *China*; 2. *Institute of Mathematics & Computer Science*, *Fuzhou University*, *Fuzhou* 350116, *China*)

Abstract: Node pair sampling in social networks is useful for friend recommendation and interest targeting. While the topology of the whole network is incomplete, or the cost of random generation of a user is very large, the performances of traditional uniform vertex sampling methods decrease quickly. So, this paper proposed a random walk based node pair sampling algorithm in large graph. Firstly, it analyzed systematically the problem of node pair sampling in social networks, and gave definitions of node pair in different hops. Secondly, it transformed the social network into an equivalent graph, whose nodes were edges in original graph, and whose edges contained two node having the same vertex. Finally, it applied random walk on the new graph and proposed a neighbor random walk sampling algorithm. The experiments show that, the proposed algorithm has less error, better performance, and is obviously better than uniform vertex sampling related methods.

Key words: graph; random walk; uniform vertex sampling; social network

0　引言

近几年，Facebook 和 Twitter 等在线社交网络在全世界都取得了巨大的成功。数以亿计的互联网用户每天使用社交网络的时间平均占他们上网时间的 22%，这远远超过了使用电子邮件的时间。与此同时，社交网络深深地改变了人们在互联网上的行为。社交网络在帮助用户维持老朋友的同时，也帮助他们结识具有共同兴趣爱好的新朋友[1]。他们为每个用户提供了在线隐私空间，并提供了聊天、短信息、e-mail、视频、语音聊天、文件共享、讨论组等多种与其他用户交互的方式。

在社交网络中，描述用户对之间的属性非常重要，可用于如下应用：

a)网络同质性检测。网络的同质性是指用户倾向于与自己有着共同偏好的用户进行链接。Singla 等人[2]通过研究发现 MSN 消息网络中存在着明显的网络同质性，即那些经常相互发送消息的用户在 Web 搜索主题、年龄和位置等个人特征上往往有着相同的偏好或特征。即使用户间从未进行过信息交互，但是他们存在着共同的好友，上述现象仍然存在。在具有同质性的网络中，可以基于用户好友的特征及偏好推断出该用户隐式的个人特征，并基于该隐式特征对用户进行有价值的推荐。

b)距离分布测量。在社会网络中，节点 A 和 B 之间的距离是它们在网络中的最短距离。网络中节点对的平均距离和网络直径是描述网络本质和进化的基本统计特征。要计算节点对的平均距离和网络的有效直径(即所有节点对距离的第 90 百分位数)，必须首先描述距离分布的特征。例如，著名的六度分割理论表明[3]：社会网络中的任意两个人都可以平均在 6 跳以内从一个人到达另一个人。六度分割理论说明人类社会是一个典型的小世界模型网络。

由于社会网络的规模往往很大(数亿个节点甚至更多)，对所有节点对的距离(如 2 跳距离)进行枚举显然是不可能的。现有的采样技术如均匀节点采样(uniform vertex sampling,

收稿日期：2014-03-20；**修回日期**：2014-05-15　　**基金项目**：福建省自然科学基金资助项目(2010J01329)

作者简介：吴春琼(1979-)，女，福建永定人，讲师，硕士，主要研究方向为数据挖掘、网络安全(fzwuchunq@163.com)；叶东毅(1964-)，男，福建福州人，教授，博士，主要研究方向为计算智能、数据挖掘.

UVS)[4]和随机游走(random walk,RW)[5]并不能直接应用于计算网络中节点的距离。对采样技术的直接应用会对估计的统计结果产生很大的偏差。例如,当在 2 跳的路径中对节点对[u,v]进行采样时,首先应用 UVS 从图 G 中选取节点 x,然后令节点 u 和 v 为 x 的两个随机邻居。随机选取两个节点 u 和 v,并且使得这两个节点包含至少一个共同的邻居非常简单。然而,下面将阐述这种采样方法并不能均匀地选取节点对。如果经采样选取节点 x,那么 x 的每个邻居节点对被选取的概率都是$\frac{2}{d_x(d_x-1)}$,其中 d_x 表示节点 x 的邻居数。令 $M(u,v)$ 表示节点 u 和 v 的公共邻居集合,那么节点对[u,v]被选取的概率为$\sum_{x\in M(u,v)}\frac{2}{d_x(d_x-1)}$,这不仅与 u 和 v 的共同邻居的个数有关,还与 u 和 v 的共同邻居的度数有关。在对节点对[u,v]进行采样的时候,由于需要访问节点 u 和 v 以及它们所有的共同邻居,因此纠正偏差的代价很大。

当未获得整个网络的拓扑,或者随机节点产生的代价很大时,UVS 方法并不适合。另外,当扫描整个社会网络拓扑时,资源受限或者代价太大时 UVS 方法也不适合。这使得对社会网络中的节点对进行统计估计变得十分困难。为了解决上述问题,本文系统地研究了大图中的节点对采样技术,提出了图拓扑信息不完全情况下的采样方法。

1　相关工作

本章介绍了大图中节点对采样的相关工作。Singla 等人[2]经过对用户的 Web 搜索主题、年龄和位置等个人信息的研究发现,MSN 消息网络有着明显的同质性。Leskovec 等人[6]通过对一系列真实的社会网络进行研究发现,网络的有效直径随着网络规模的增长而逐渐减小,这与现有的网络进化模型的基本假设正好相反。现有的图采样研究工作主要集中在设计有效并准确的图特征度量方法上,如节点的度分布和节点团体的拓扑结构。这些采样方法已经被广泛应用于描述 P2P 网络和在线社交网络等复杂网络的特征上。

现有的图采样工作主要有宽度优先搜索(breadth-first-search,BFS)[7]、随机游走(RW)[5]、均匀顶点采样(UVS)[4]和 Metropolis-Hasting 随机游走(MHRW)[8]。BFS 方法在采样中更容易倾向于那些图中未知度数大的节点,因而采样结果存在着偏差。RW 方法同样倾向于那些度数大的节点,然而该偏差是可知的,并且可以被纠正。与均匀顶点采样方法相比,RW 对度数大的节点的估计误差较小,当 UVS 的代价较大时,这种现象更为明显。为了对 RW 方法进行纠偏,可以对采样得到的值重新加权,从而得到图特征的无偏估计。与这种纠偏式的 RW 相比,MHRW 通过 Metropolis-Hasting 技术对随机游走过程进行修改,其目标在于均匀地对节点进行采样。文献[9]对比了 RW 和 MHRW 两种方法的准确性,大量的实验表明,纠偏方法得到的 RW 的准确性始终优于或等价于 MHRW 算法。在 RW 算法中,纠偏所占的混合时间决定了采样算法的性能。对于大多数社交网络而言,纠偏式 RW 方法所用的时间远远大于预期的时间。于是,出现了大量减少纠偏时的混合时间的方法[10]。与这些研究不同的是,本文对网络中带约束的节点对采样问题进行了合理的理论分析。

2　问题描述

令 $G=(V,E)$ 表示一个无向图,其中 V 表示节点的集合,E 表示边的集合。令 $u,v\in V$ 表示节点,$(u,v)\in E$ 表示边,[u,v]表示 u 和 v 两个节点组成的节点对,图 G 中的边无自环,即 $\forall u,(u,u)\notin E$。$(u,v)\neq(v,u)$,并且$[u,v]\neq[v,u]$。本文通过采样方法测量如下定义的节点对特征。

定义 1　全集 $S=\{[u,v]:u,v\in V$ 并且 $u\neq v\}$。

定义 2　1 跳子集 $S^{(1)}=\{[u,v]:(u,v)\in E\}$。

定义 3　2 跳子集$\{S^{(2)}=[u,v]:\exists x\in V,(u,x)\in E\wedge(v,x)\in E,u\neq v,u,v\in V\}$。

定义 4　1 到 2 跳子集 $S^{(2+)}=S^{(1)}\cup S^{(2)}$。

很容易发现 $S^{(1)}$ 和 $S^{(2)}$ 的交集是非空的,即网络中存在着三角形。如果 G 中的边(u,v)的两个顶点 u 和 v 没有共同的邻居,那么有$[u,v]\in S^{(1)}$并且$[u,v]\notin S^{(2)}$,即 $S^{(2)}$ 可能并不包含 $S^{(1)}$ 中的所有元素。$S^{(1)}$ 包含图中所有距离为 1 的节点对,$S^{(2+)}$ 包含图中所有距离不大于 2 的节点对。

定义函数 $F:V\times V\to\mathbb{R}$。对每个节点对$[u,v]$,$F(u,v)$为该节点对在某指定属性下的值,如节点 u 和 v 的公共邻居个数。$F(u,v)$的值可能与 $F(v,u)$的值不相等,如当定义 $F(u,v)$为节点 u 的个数与节点 u 和 v 的公共邻居个数的差值时。令$\{a_1,\cdots,a_K\}$为 $F(u,v)$的值域。本文通过采样方法来估计节点对的分布 $\omega=(\omega_1,\cdots,\omega_K)$,$\omega^{(1)}=(\omega_1^{(1)},\cdots,\omega_K^{(1)})$,$\omega^{(2)}=(\omega_1^{(2)},\cdots,\omega_K^{(2)})$和 $\omega^{(2+)}=(\omega_1^{(2+)},\cdots,\omega_K^{(2+)})$。其中 ω、$\omega^{(1)}$、$\omega^{(2)}$和 $\omega^{(2+)}(1\leqslant k\leqslant K)$为集合 S、$S^{(1)}$、$S^{(2)}$和 $S^{(2+)}$中节点对[u,v]的个数与 $F(u,v)=a_k$ 的比值。

定义 5　$S^{(1-)}=S^{(1)}\backslash S^{(2)}$,即对于 $S^{(1-)}$中的每个元素[u,v]$\in S^{(1-)}$,u 和 v 是相链接的,但是却没有公共的邻居。

定义 6　$\omega^{(1-)}=(\omega_1^{(1-)},\cdots,\omega_K^{(1-)})$,其中 $\omega_k^{(1-)}(1\leqslant k\leqslant K)$是 $S^{(1-)}$中节点对[u,v]的个数与 $F(u,v)=a_k$ 的比值。

令 $\alpha=\frac{|S^{(1-)}|}{|S^{(1)}|}$和$\beta=\frac{|S^{(1)}|}{|S^{(2)}|}$,可得

$$\omega_k^{(2+)}=\frac{|S^{(1-)}|\omega_k^{(1-)}+|S^{(2)}|\omega_k^{(2)}}{|S^{(1-)}|+|S^{(2)}|}=\frac{\alpha\omega_k^{(1-)}+\beta\omega_k^{(2)}}{\alpha\beta+1}\quad(1)$$

在大多数社交网络中,由于 α 和 β 的值非常小,$\omega_k^{(2+)}$ 与 $\omega_k^{(2)}$ 非常接近。由于$|S^{(1)}|=2|E|$,$|S|=|V|(|V|-1)$和$|S^{(2)}|^2$的值都远远大于$|V|$,因此在对节点规模以亿计的网络中计算 ω 和 $\omega^{(2+)}$时,采样估计是最好的选择。

3　基于随机游走的节点对采样方法

在连通图 G 中,如果 G 的拓扑结构部分可知,或者产生随机 ID 的代价很高,那么不适合采用 UVS 方法进行采样。在这种情况下,本文应用随机游走模型进行节点对的采样分析。随机游走的基本思想是从某个初始点出发,从当前节点的邻居节点中随机选取一个节点作为当前节点的下一跳节点。随机游走到达该邻居节点,并对节点的相关信息进行采样。令随机游动的稳态分布为 $\pi=(\pi_v:v\in V)(\pi_v=d_v/2|E|)$,在一个连通的非二部图中,随机游走到节点 $v\in V$ 的概率收敛于 π_v。下述的顶点采样算法是非均匀的采样算法:在每一步,根据概率分布 π 从 V 选出一个节点。随机游走方法在采样时更倾向于那

些度数大的节点,该偏差是可以被纠正的[11,12]。与均匀顶点采样方法相比,随机游走方法对度数大的节点的估计误差相对较小。

3.1 S 和 $S^{(1)}$ 中的节点对采样

在图 G 中,分别进行两个独立的随机游走对节点采样,在第 i 步,两个随机游走采集到的节点分别记为 u_i 和 v_i,得到的节点对为 $\{[u_i,v_i]\}_{i=1,\cdots,n}$。这种随机游走可以看做图 $G^{(2)}=(V^{(2)},E^{(2)})$ 上的正规随机游走,其中 $V^{(2)}=\{[u,v]:u,v\in V\}$,$E^{(2)}=\{([u,v],[x,y]):(u,x),(v,y)\in E\}$。在图 $G^{(2)}$ 中,节点 $[u,v]$ 有 d_ud_v 个邻居。当图 G 是连通的非二部图时,图 $G^{(2)}$ 也是连通的非二部图,于是图 $G^{(2)}$ 上随机游走的稳态概率 $\pi_S=(\pi_{[u,v]}:u,v\in V)$ 为

$$\pi_{[u,v]}=\frac{d_ud_v}{4|E|^2}\qquad u,v\in V \tag{2}$$

由于随机游走可能通过稳态分布 $\sum_{u\in V}\pi_{[u,v]}=\frac{\sum_{u\in V}d_u^2}{4|E|^2}$ 进行采样,$\omega_k(1\leqslant k\leqslant K)$ 可通过式(3)进行估计。

$$\bar{\omega}_k^*=\frac{1}{J}\sum_{i=1}^{n}\frac{1(F(u_i,v_i)=a_k)1(u_i\neq v_i)}{d_{u_i}d_{v_i}} \tag{3}$$

其中:$J=\sum_{i=1}^{n}\frac{1(u_i\neq v_i)}{d_{u_i}d_{v_i}}$。

定理 1[13] 当图 G 是一个连通的非二部图时,$\bar{\omega}_k^*$ 是 ω_k 的一个渐进的无偏估计器。

为了估计 $S^{(1)}$ 中节点对的特征,基于图 G 上的随机游走采样节点对 $\{[u_i,v_i]\}_{i=1,\cdots,n}$,其中节点 u_i 和 v_i 分别是随机游走在第 i 和 $i+1$ 步采样得到的节点。显而易见,$(u_i,v_i)\in E$ 是图 G 中的一条边。当随机游走达到稳态时,采样得到的边的概率是相等的[14],于是 $\omega_k^{(1)}(1\leqslant k\leqslant K)$ 可通过式(4)估计得到。

$$\bar{\omega}_k^{(1*)}=\frac{1}{n}\sum_{i=1}^{n}1(F(u_i,v_i)=a_k) \tag{4}$$

定理 2[13] 当图 G 是一个连通的非二部图时,$\bar{\omega}_k^{(1*)}(1\leqslant k\leqslant K)$ 是 $\omega_k^{(1)}$ 的一个渐进的无偏估计器。

3.2 $S^{(2)}$ 中的节点对采样

在从 $S^{(2)}$ 中进行随机采样时,本文提出了一种邻居随机游走(neighborhood random walk, NRW)节点对采样方法。令图 $G'=(V',E')$,节点的集合 $V'=\{(u,v):(u,v)\in E\}$,边的集合 $E'=\{((u,v),(u,v')):(u,v)\in E,(u,v')\in E,v\neq v'\}$,邻居随机游走节点对采样方法可以看做图 G' 上正规的随机游走。令 G 中的边 (u,v) 为 NRW 在图 G' 中的初始节点,$N_{(u,v)}$ 为 G 中与节点 u 或 v 相链接的且不含 (u,v) 的边集,于是有 $|N_{(u,v)}|=d_u+d_v-2$。NRW 从 $N_{(u,v)}$ 中随机选取一条边作为下一个采样边,该过程可以建模成一个传输矩阵为 $P^{\mathrm{NRW}}=[P^{\mathrm{NRW}}_{e,e'}]$ 的马尔可夫链,其中 $e=(u,v)$ 和 $e'=(u',v')$ 为 E 中的边,$P^{\mathrm{NRW}}_{e,e'}$ 表示 NRW 模型中当前边为 e 的情况下下一跳的边为 e' 的概率。

$$P^{\mathrm{NRW}}_{(u,v),(u',v')}=\begin{cases}\dfrac{1}{d_u+d_v-2} & (u,v)\in E\text{ 且 }(u',v')\in N_{(u,v)}\\ 0 & \text{否则}\end{cases} \tag{5}$$

在图 G' 中,节点 (u,v)(图 G 中的边)的邻居节点个数为 d_u+d_v-1,因此度数为 d_u+d_v-1。此外,图 G' 有 $|E'|=M/2$ 条边,其中 $M=\sum_{y\in V}d_y(d_y-1)$。

定理 3 当图 G 是一个连通的非二部图时,NRW 的稳态分布为 $\pi_E=(\pi_{(u,v)}:(u,v)\in E)$,其中 $\pi_{(u,v)}$ 的计算为

$$\pi_{(u,v)}=\frac{d_u+d_v-2}{M}\qquad (u,v)\in E \tag{6}$$

证明 假设 NRW 当前到达边 (u,v) 的概率为 $\pi_{(u,v)}$,那么 NRW 在下一步选择边 (u',v') 的概率 $P_{(u',v')}$ 为

$$P_{(u',v')}=\sum_{(u,v)\in N_{(u',v')}}\pi_{(u,v)}P^{\mathrm{NRW}}_{(u,v),(u',v')}=\frac{|N_{(u',v')}|}{M}=\pi_{(u',v')}$$

于是,π_E 是传输矩阵为 P^{NRW} 的马尔可夫链的稳态分布。由于图 G 是连通的非二部图,图 G' 也是连通的非二部图。因为 G' 中的节点 (u,v) 的度数为 d_u+d_v-1,并且 NRW 可以看做 G' 上正规的随机游走,所以 NRW 到达边 $(u,v)\in E$ 的概率收敛于 π_E。

NRW 可以看做是 G' 上正规的随机游走,它用相同的概率随机地从图 G' 上对边进行采样,因此节点对 (u,v) 被采集到的概率是 $m(u,v)/M$。基于采集到的节点对 $\{[u_i,v_i]\}_{i=1,\cdots,n}$,$\omega_k^{(2)}(1\leqslant k\leqslant K)$ 的估计公式为

$$\pi_{(u,v)}=\frac{d_u+d_v-2}{M}\qquad (u,v)\in E \tag{7}$$

其中:$H=\sum_{i=1}^{n}\frac{1}{m(u_i,v_i)}$。

定理 4[13] 当图 G 是一个连通的非二部图时,$\bar{\omega}_k^{(2*)}(1\leqslant k\leqslant K)$ 是 $\omega_k^{(2)}$ 的一个渐进的无偏估计器。

4 实验设计与结果分析

4.1 数据集与评价标准

本文通过真实的社会网络数据集对提出的算法进行验证,数据集的基本信息如表 1 所示。Wikipedia 是用户通过协同的方式编写的自由百科全书。在 Wikipedia 数据集中,网络的节点是用户,网络的有向边表示一个用户对另一个用户进行了提议。Gnutella 是一个 P2P 文件共享网络,网络的节点是主机,网络的边表示主机之间存在着连接。在实验中,将数据集的有向边转换为无向边,并应用规范均方差(normalized mean square error,NMSE)来评价算法的性能。

$$\mathrm{NMSE}(\bar{\omega}_k)=\frac{\sqrt{E[(\bar{\omega}_k-\omega_k)^2]}}{\omega_k}\qquad k=1,2,\cdots \tag{8}$$

表 1 数据集基本信息

社会网络	完全图		最大连通子图	
	节点个数	边的个数	节点个数	边的个数
Wikipedia[14]	7 115	103 689	7 066	103 663
Gnutella[15]	10 876	39 994	10 876	39 994

实验应用 NMSE 来评价估计值 $\bar{\omega}_k$ 与真实值 ω_k 之间的相对误差。由于采用的是相对误差,当 $\bar{\omega}_k$ 取值很小时,$\bar{\omega}_k$ 的 NMSE 大到 1 时仍然是可以接受的。

4.2 实验结果

本实验将独立加权顶点采样方法记为 IWVS,将基于 Metropolis-Hasting 方法的顶点采样方法记为 MHWVS,并将本文提出的邻居随机游走方法记为 RW。

首先,在 Wikipedia 数据集中,为了测试算法在不完全数据集下的性能,分别取 0.5% 和 1% 的 $S^{(1)}$ 数据。三种算法的在 $\bar{\omega}_k^{(1)}$ 下的 NMSE 对比分别如图 1 和 2 所示。从这两个图中可以看出,随着节点互为邻居数目的增加,三种算法的 NMSE 的

收敛情况几乎是一致的。

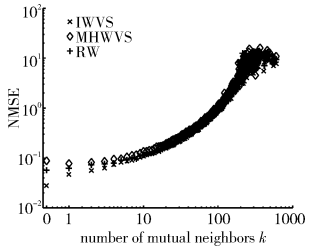

图 1　Wikipedia 数据集下 NMSE 随互为邻居数的变化($0.5\%\,|S^{(1)}|,\overline{\omega}_k^{(1)}$)

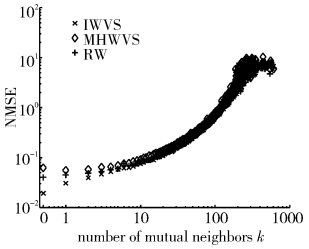

图 2　Wikipedia 数据集下 NMSE 随互为邻居数的变化($1\%\,|S^{(1)}|,\overline{\omega}_k^{(1)}$)

然后,在 Gnutella 数据集中,分别取 0.5% 和 1% 的 $S^{(1)}$ 数据。三种算法在$\overline{\omega}_k^{(2)}$ 下的 NMSE 对比分别如图 3 和 4 所示。从图中可以看出,Gnutella 数据集在$\overline{\omega}_k^{(2)}$ 下与 Wikipedia 数据集在$\overline{\omega}_k^{(1)}$ 下有着相同的结果,这表明算法不依赖于具体的数据集,并且在$\overline{\omega}_k^{(1)}$ 和$\overline{\omega}_k^{(2)}$ 下有着相同的结论。

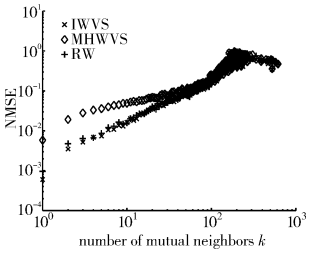

图 3　Gnutella 数据集下算法的 NMSE 对比($0.5\%\,|S^{(2)}|,\overline{\omega}_k^{(2)}$)

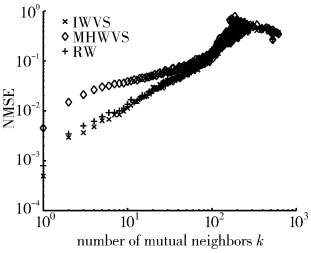

图 4　Gnutella 数据集下算法的 NMSE 对比($1\%\,|S^{(2)}|,\overline{\omega}_k^{(2)}$)

接下来,测试了三种算法在不同数据集下随着共同邻居数的增加算法的误差对比。在本实验中,两个数据集都取 1% 的 $S^{(2)}$ 数据,实验结果如图 5 和 6 所示。图中表明,当网络中的最大连通子图密集时,各种算法的误差几乎是相同的。

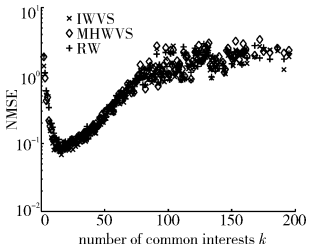

图 5　Wikipedia 数据集下 NMSE 随共同邻居数的变化($1\%\,|S^{(2)}|,\overline{\omega}_k^{(2)}$)

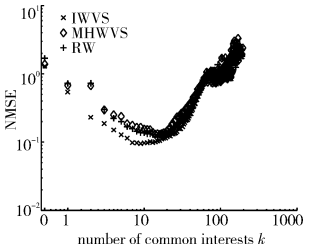

图 6　Gnutella 数据集下 NMSE 随共同邻居数的变化($1\%\,|S^{(2)}|,\overline{\omega}_k^{(2)}$)

最后,对算法的执行效率进行了对比。本实验在两个数据集下取数据集的 1% 数据,测试了三种算法的 NMSE。令共同邻居数为 $k=20$,实验结果如图 7 所示。从图 7 可以看出,本文提出的邻居随机游走算法在两种数据集下的执行时间是最短的。

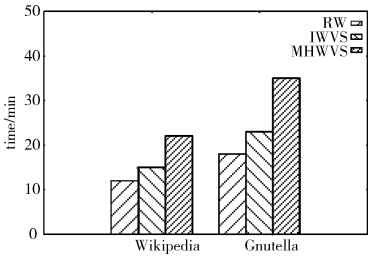

图 7　算法的执行性能对比

5　结束语

节点对采样是社会网络分析中的基础研究内容,可广泛应用于网络的同质性测量、网络的直径分析等。本文系统地研究了集合 S、$S^{(1)}$ 和 $S^{(2)}$ 下节点对的统计特征。提出了一种邻居随机游动方法,该方法既可以用于网络拓扑已知的网络,也可以用于拓扑结构不完全的有向非二部图网络。实验结果表明,本文提出的方法统计误差小,并且执行效率高,明显优于均匀采样的相关算法。

参考文献:

[1] 马帅, 曹洋, 沃天宇, 等. 社会网络与图匹配查询[J]. 中国计算机学会通讯, 2012, 8(4): 20-24.

[2] SINGLA P, RICHARDSON M. Yes, there is a correlation: from social networks to personal behavior on the Web[C]//Proc of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 655-664.

[3] 余学军. 六度分割理论成就 SNS[J]. 信息网络, 2009(11): 37-37.

[4] XUESONG L, STÉPHANE B. Sampling connected induced subgraphs uniformly at random[C]//Proc of the 24th International Conference on Scientific and Statistical Database Management. 2013: 781-792.

[5] 蔡君, 余顺争. 基于随机聚类采样算法的复杂网络社团探测[J]. 计算机应用研究, 2013, 30(12): 3560-3563.

[6] LESKOVEC J, HORVITZ E. Planetary-scale views on a large instant-messaging network[C]//Proc of WWW. 2008: 915-924.

[7] ZHOU Rong, HANSEN E A. Breadth-first heuristic search[J]. Artificial Intelligence, 2006, 170(4): 385-408.

[8] GJOKA M, KURANT M, BUTTS C T, *et al*. Walking in facebook: a case study of unbiased sampling of OSNs[C]//Proc of IEEE INFOCOM. 2010: 2498-2506.

[9] 夏放怀, 沈振康, 唐朝京, 等. 一种用于实时体绘制系统的自适应采样算法[J]. 电子学报, 2002, 30(3): 367-371.

[10] MOHAISEN A, YUN A, KIM Y. Measuring the mixing time of social graphs[C]//Proc of ACM SIGCOMM Internet Measurement Conference. 2010: 390-403.

[11] HECKATHORN D D. Respondent-driven sampling Ⅱ: deriving valid population estimates from chain-referral samples of hidden populations[J]. Social Problems, 2002, 49(1): 11-34.

[12] SALGANIK J, HECKATHORN D D. Sampling and estimation in hidden populations using respondent-driven sampling[J]. Sociological Methodology, 2004, 34(1): 193-239.

[13] WANG Ping-hui, ZHAO Jun-zhou, LIU J C S, *et al*. Sampling node pairs over graphs[EB/OL]. (2012). http://www.cse.cuhk.edu.hk/%7ephwang/samplingpairreport.pdf.

[14] LESKOVEC J, HUTTENLOCHER D, KLEINBERG J. Predicting positive and negative links in online social networks[C]//Proc of WWW. 2010: 641-650.

[15] RIPEANU M, FOSTER I T, IAMNITCHI A. Mapping the Gnutella network: properties of large-scale peer-to-peer systems and implications for system design[J]. IEEE Internet Computing Journal, 2002, 6(1): 50-57.