在线社交网络的 UNI64 采样方法①

许南山, 李 浩, 卢 罡

(北京化工大学 信息科学与技术学院, 北京 100029)

摘 要: 在对社交网络采样方法进行研究时, 常以拒绝-接受采样法得到的样本作为对照来评价其他采样方法的优劣. 由于各种在线社交网络陆续将其用户 ID 系统由 32 位升级为 64 位, 导致拒绝-接受采样法的采样命中率近乎为零. 本文根据在线社交网络的特点, 以新浪微博为例, 对其用户 ID 分布情况进行分析, 提出了一种改进的拒绝-接受采样法 UNI64. 该方法通过分析网络有效 ID 样本的分布情况, 结合聚类的方法将整个样本空间划分为有效区间和无效区间, 并使采样算法避开无效区间, 仅在有效区间内生成待测样本, 从而有效提高了拒绝-接受采样法在有效样本极为稀疏的样本空间内采样的命中率.

关键词: 在线社交网络; 采样方法; 随机采样; 新浪微博; 层次聚类

UNI64 Sampling Method on Online Social Networks

XU Nan-Shan, LI Hao, LU Gang

(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: When studying the sampling methods on online social networks, samples collected by acceptance-rejection method are usually used as the “ground truth” to estimate the pros and cons of other sampling methods. The acceptance rate of the original acceptance-rejection method slumps dramatically since OSN sites updated their user ID from 32bit to 64bit. According to the characteristics of online social networks and taking Sina Weibo for example, we analyzed the distribution of user IDs in Sina Weibo, and proposed an improved acceptance-rejection method called UNI64. In this method, the user ID space is divided into valid intervals and vacant intervals by analyzing the distribution of valid sample IDs and utilizing cluster method. The sampling method generates candidate IDs only in valid intervals, so that the acceptance rate could be effectively improved even in a sparse-distributed user ID space.

Key words: online social networks; sampling method; random sampling; Sina Weibo; hierarchical cluster

1 引言

在线社交网络(Online Social Networks, OSN)作为 Web 2.0 时代网络应用的杰出代表, 在近十年时间内迅速蔓延全球. 以 Facebook、Twitter 为代表的新型社交网络的出现使在线社交网络用户数量呈现爆炸型增长. 据统计全球 Facebook 注册用户已达 14 亿. 而在国内, 新浪微博的注册用户也已超过 5 亿. 在线社交网络正在以其独有的特点对现实社会产生着广泛而深入的影响. 同时, 在线社交网络近年来已吸引了一大批学者对其特点进行分析和研究, 并且已经在网络结构演化、用户行为分析, 以及信息传播模型研究等方面取得了一定的成果[1-4].

绝大多数对在线社交网络的实证性研究都要以真实的网络节点和结构数据为基础[5-7], 然而获取这些数据并不容易. 出于商业机密、用户隐私等原因, 在线社交网络的服务提供商不可能将其全网数据提供出来用于科学研究; 加之数据量庞大并且随时间不断变化, 直接使用全网数据进行研究也并不可行. 因此, 得到一个能够代表真实网络某个或某些方面特征的网络样本数据集合成为在线社交网络研究的必要条件.

① 基金项目:北京高等学校青年英才计划(YETP0506)
收稿时间:2014-04-10;收到修改稿时间:2014-05-09

为了得到具有代表性的样本数据, 研究人员提出了很多针对网络的采样方法. 这些方法基本可以分为两类: 基于图遍历的方法和基于随机游走的方法[8]. 常见的基于图遍历的采样方法有 Bread-First Search (BFS)、Depth-First Search (DFS)、Forest Fire (FF)、SnowBall Sampling (SBS)等, 常见的基于随机游走的算法有 Random Walk (RW)、Re-Weighted Random Walk (RWRW)和 Metroplis-Hasting Random Walk (MHRW)等. 然而, 由于缺少全网数据进行验证, 通过这些方法得到样本的代表性就无法直接进行评判.

美国加州大学欧文分校的 Gjoka 等人在对 Facebook 进行研究时, 提出了一种基于接受-拒绝采样方法的无偏均匀采样法, 并将该方法采集的样本作为其他采样方法的评估基准值[9]. 该方法被其命名为 UNI 方法. 然而他们也明确指出, 该方法仅适用于采集用户 ID 为 32 位整数的网络系统. 但是现在多数在线社交网络的用户 ID 都已经升级为 64 位整数, 这就使得 UNI 方法的采样命中率急剧下降, 接近于零.

本文通过对 UNI 方法进行分析, 针对该方法采样命中率在 64 位 ID 系统中急剧下降的问题提出了一套改进的 UNI 方法, 我们将其称之为 UNI64 方法. 并且, 我们以新浪微博为实例, 对其用户 ID 分布情况进行了分析和研究. 最后在实际网络中对本文提出的算法的有效性进行了验证.

2 UNI方法分析

UNI 方法是基于接受-拒绝采样, 针对用户 ID 的均匀随机采样方法. 其最初被提出时是被用于对 Facebook 的用户 ID 进行均匀采样. 该方法主要由两步组成: (1)从系统所有可能的 ID 范围内随机生成一个整数值作为待测 ID, (2)在真实在线社交网络中查询, 若有用户与该 ID 对应, 则称其为有效 ID, 接受并作为样本值保存; 否则拒绝并抛弃该 ID. 循环执行这两步直到满足停止条件.

通过对 UNI 方法的分析, 不难看出该方法的效率对于随机采样的命中率是极其敏感的. 采样命中率越低, 获得一定数量有效 ID 所需的采样次数就越多. Gjoka 等人在提出 UNI 方法时, Facebook 的用户 ID 系统还是 32 位, 那时 Facebook 用户总量约为 2×10^8, 则随机采样得到一个有效 ID 的概率为 1/22[9]. 然而随着越来越多的用户使用在线社交网络, Facebook、Twitter、微博等热门在线社交网络都已将其用户 ID 升级为 64 位, 在 2^{64} 整数空间内直接使用 UNI 方法的命中率太低, 已经无法在实际采样中应用. 以新浪微博为例, 尽管其注册用户已经超过 5 亿, 但如果在 64 位用户 ID 所能表示的$[0,2^{64}-1]$范围内进行随机采样, 其理论采样命中率也是微乎其微.

3 UNI64采样方法

3.1 UNI64 方法概述

首先, 我们引入理论采样命中率和实际采样命中率的概念. 其中实际采样命中率定义为采样获得的有效 ID 个数与采样次数的比值, 即:

$$\text{实际采样命中率}=\frac{\text{采样有效ID个数}}{\text{采样次数}}$$

并将理论采样命中率定义为 ID 系统中全部有效 ID 个数与该系统所能表示的所有 ID 个数的比值, 即:

$$\text{理论采样命中率}=\frac{\text{系统中有效ID个数}}{\text{系统中所有ID个数}}$$

在进行多次随机采样实验后实际采样命中率是应该收敛于理论采样命中率的, 因此我们可以通过理论采样命中率来估计随机采样的实际命中率.

提高 UNI 方法的实际采样命中率是使其能够应用于 64 位 ID 系统中的关键. 为此我们提出一个假设:

考虑到在线社交网络为了便于对用户 ID 系统进行管理, 不太可能是在整个系统中随机分配 ID, 同时由于用户 ID 从初始的 32 位升级为目前的 64 位, 因此我们推断有效 ID 在系统中的分布不是随机分散于整个 64 位正整数空间的, 而是呈现非均匀的分布.

64 位 ID 系统能够表示的区间非常大, 如果有效 ID 在一些区域分布比较密集, 则必然会在另一些区域出现大量的“无人区”. 若能够识别出这些分布密度大的区域, 并控制 UNI 方法避开“无人区”, 仅在有效区域进行采样, 那么算法的实际命中率就应该能够得到很大提高.

根据这个思路, 我们设计了如下工作: 首先, 以新浪微博为例, 随机采集一定量的有效 ID, 并分析这些有效 ID 在系统中的分布情况; 其次设计算法使其能够根据一定数量有效 ID 样本的分布, 从整个系统中划分出有效区间; 最后修改 UNI 方法, 使其能够针对 64 位的用户 ID 进行随机采样, 并在新浪微博和 Facebook 上进行实验, 对该方法进行评估.

我们将这一系列用以提高 UNI 算法在 64 位 ID 系统中实际采样命中率的方法称为“UNI64 采样方法”.

3.2 新浪微博用户 ID 分布情况分析

我们通过调用新浪微博提供的 API 接口获取了 96645440 个用户 ID 值. 为了验证之前关于有效用户 ID 在整个用户 ID 空间中分布的假设, 我们对这 9 千多万条数据的分布情况进行了两种统计: 第一种是按照 ID 值的数值位数进行统计, 从而对整个样本数据的分布进行一个直观的了解; 第二种是查看用户 ID 值在指定区间之间的分布情况, 从而可以针对某个区间进行详细划分, 了解更细致的分布情况.

由于 64 位正整数的范围是[0, 2^{64}-1], 故最大的用户 ID 值不超过 20 位, 则根据 ID 值的数值位数进行统计后得到表 1 所示结果. 由于各长度 ID 的数量相差较为悬殊, 故将 ID 数量取对数后进行统计, 得到分布图, 如图 1 所示.

表 1　不同长度 ID 值数量统计结果

ID 长度	ID 数量	ID 长度	ID 数量
1	1	11	3
2	2	12	0
3	0	13	0
4	1	14	0
5	655	15	0
6	1588	16	0
7	1293	17	0
8	17727	18	0
9	75045	19	0
10	96549124	20	0

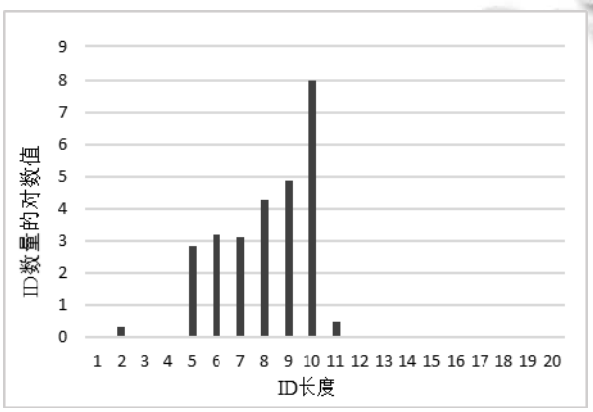

图 1　新浪微博用户 ID 长度频率分布图

在样本数据中, 超过 99%的 ID 值的长度都是 10 位. 这说明新浪微博用户 ID 基本集中在[1×10^9, 1×10^{10})之间.

接着对长度为 10 的有效 ID 样本进一步分析, 将[1×10^9, 1×10^{10})区间以 1×10^9 为单位进行等距划分, 然后统计长度为 10 的 ID 样本在其中的分布情况. 统计结果如表 2 所示.

表 2　长度为 10 的 ID 样本在均分区间内的统计结果

区间范围	ID 个数
[1×10^9, 2×10^9)	22138453
[2×10^9, 3×10^9)	46362133
[3×10^9, 4×10^9)	28048531
[4×10^9, 5×10^9)	7
[5×10^9, 6×10^9)	0
[6×10^9, 7×10^9)	0
[7×10^9, 8×10^9)	0
[8×10^9, 9×10^9)	0
[9×10^9, 10×10^9)	0
总数	96549124

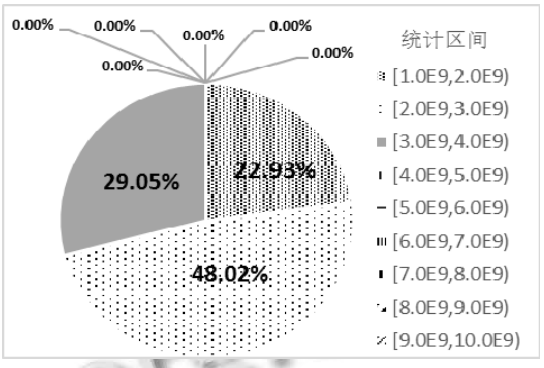

图 2　长度为 10 的 ID 样本在均匀区间的分布比例

表 2 显示长度为 10 的样本 ID 基本都分布在[1×10^9, 2×10^9), [2×10^9, 3×10^9), [3×10^9, 4×10^9)这三个区间中. 分析表 2 中区间分布的比例, 见图 2, 可以看出接近一半都分布在[2×10^9, 3×10^9)区间中.

3.3 有效区间划分算法

通过前面的分析, 可以看到新浪微博有效 ID 的分布非常集中, 几乎 99%以上样本都分布在[1×10^9, 4×10^9)之间, 这是由于新浪微博的用户 ID 系统也是从 32 位升级来的, 因此只要在这个区间内进行采样, 命中率就很容易达到要求. 然而, 当遇到有效 ID 的分布比较复杂时, 这种通过单纯统计样本数据分布决定有效区间的方法就行不通了. 因此需要提出一种通用的自适应算法, 使其能够自动对样本数据进行分析, 然后输出一组能够满足一定实际采样命中率的有效区

间.

由于有效 ID 样本有序分布在一维空间中, 算法需要根据样本之间的距离和目标命中率将其划分在不同的区间内且满足特定条件, 因此该问题可抽象为一维数轴的区间划分问题. 问题描述如下:

表 3 符号表

符号	定义
X	升序排列的有效 ID 样本集合
N	X 中的元素个数, 即有效 ID 样本个数
D	一个有序集合, 其元素是正整数且按升序排列
P	目标命中率, 即期望达到的实际命中率
$\hat{P}$	某区间估计命中率
T	系统中有效 ID 总数
$\hat{\theta}$	某区间内有效 ID 个数的估计值
dia	有效直径
S	一个区间内相邻样本的平均距离
$f=<D, S>$	有序集合 D 与其元素间平均距离的二元组

定义 1. 有效直径 dia 是以一个有效样本 ID 值为中心的数轴区间长度, 满足 $\frac{\hat{\theta}}{dia}=P$. 有效直径是保证区间内估计采样命中率不小于目标命中率的最大区间长度的理论值. 其中, 区间内有效 ID 个数的估计值可由 $\hat{\theta}=\frac{T}{n}$ 求得.

问题描述:

设: (1)$D=<a_1,a_2,a_3,\ldots,a_k>$, $a_i\in Z^+$且 $a_i < a_{i+1}$.

(2)F 为 f 的集合, $F=<f_1,f_2,f_3,\ldots,f_m>$, 其中 $f_i=<D_i,S_i>$, $i\in[1,m]$, 且满足 $\forall b\in D_i$, $\forall c\in D_{i+1}$, 有 $b<c$ 且 $\Phi=\bigcap_1^m D_i$

已知: $X=<x_1,x_2,x_3,\ldots,x_n>$, 有效直径 dia

求: 集合 $F=<f_1,f_2,f_3,\ldots,f_m>$, 满足:

(1) $X=\bigcup_1^m D_i$

(2) $\|F\|$ 最小

(3) 且对于 $\forall f_i\in F$, 有 $S_i=\frac{a_{\max}-a_{\min}}{\|D_i\|-1}\le dia$

在计算之前首先需要人为给定目标命中率 P. 区间估计命中率 $\hat{p}$ 可由区间有效 ID 估计值和区间大小求得, 算法要在保证区间估计命中率 $\hat{p}$ 达到目标命中率 P 的前提下使划分出来的有效区间数量尽可能地少, 因此可以利用自顶向下层次聚类的思想, 迭代地对数轴区间进行合并, 直到在满足命中率 P 的前提下没有区间可以合并为止. 具体算法如下:

算法 1:

输入: 样本数据 $X=<x_1,x_2,x_3,\ldots,x_n>$,目标采样命中率为 P, 总体数 T

输出: 集合 F

开始:

1)计算有效区间内有效 ID 个数估计值 $\hat{\theta}=\frac{T}{n}$, 有效直径 $dia=\frac{\hat{\theta}}{P}$

步骤 1:

/*顺序遍历 X, 合并平均距离小于 dia 的元素*/

2)$i = 1$; $j = 2$; $k = 1$; *new link list F; new node* $f_k=<D,S,next>$

3)while($j <= n$)

4) if ($(x_j - x_i)/(j - i) \le dia$) //判断合并新元素后平均距离是否超过有效半径长度

5) add x_j to the tail of $f_k.D$ //将新元素加入当前集合尾部

6) j++ //按顺序向后

7) else //当合并新元素后平均距离超过有效半径时完成当前合并

8) $f_k.S=\frac{x_j-x_i}{\|f_k.D\|-1}$ //计算当前集合的平均距离

9) add f_k to the tail of F //当前 f_k 构造结束, 并将其加入 F 尾部

10) k++; *new node* $f_k=<D,S,next>$//创建一个新 f

11) $i=j+1$; $j=j+2$

12) End if

13)End while

步骤 2:

/*合并满足条件的相邻节点 f_k */

14)*new set V=null* //V 是存放不满足合并条件的 f 的集合

15)while ((F-V)≠*null*)

16) find f_k with the smallest S from(F-V)

17) $M = f_k.D\cup f_{k+1}.D$ //将相邻集合合并后的结果暂时存放在集合 M 中, 当判断满足条件后再将它们真正合并

18) if ($S_M=\frac{a_{\max}-a_{\min}}{\|M\|-1}\le dia$) //判断两集合合并后元素平均距离是否超过有效半径

19) $f_k.D = M$//将合并后的结果存放在前一个 f 中

20) $f_k.S = S_M$

21) $f_k.next = f_{k+1}.next$

22) remove f_{k+1} from F //由于已将后一个节点的内容合并入前一个节点, 故将有一个节点移除

23) clear V

24) else

25) add f_k to V

26) End if

27)End while

算法步骤 1 的功能是将顺序排列的元素合并成为平均距离满足要求的基本集合并将其存储在二元结构 f 中以备步骤 2 使用; 步骤 2 是将合并后平均距离仍能满足要求的下标相邻的 f 中集合进行合并, 以减少集合的数量. 在步骤 2 中, 按集合的平均距离从小到大的顺序选出候选合并对象, 因为平均距离越小的集合能够合并与它距离越远的集合, 也就意味着它合并其他集合的能力越强. 若选出的集合与其相邻集合合并后的平均距离大于有效半径, 说明该集合不具有合并能力, 则将其加入集合 V 中, 并从(F-V)集合中选出平均距离最小的, 并尝试与它下标相邻的集合进行合并. 当没有任何可以合并的集合时, 即集合 F 中的元素与 V 集合中的元素相等时, 算法结束.

算法步骤 1 经过一次遍历就能得到结果, 若输入样本个数为 n, 则步骤 1 的时间复杂度为 $O(n)$. 若步骤 1 得到的集合 F 中有 m 个元素, 由于步骤 2 需要对所有集合的平均距离进行排序, 使用快速排序的时间复杂度为 $O(m \log m)$; 合并操作最小时间复杂度为 $O(m)$, 最坏情况下的时间复杂度为 $O(m!)$.

3.4 有效区间内随机采样

在有效区间内进行随机采样时, 为了保证采样结果的密度分布符合有效 ID 的实际密度分布, 需要对每个有效区间内的采样次数进行控制. 在总采样次数一定的情况下, 分配在每个区间内的采样次数应该和区间的大小成正比. 又由定义 1 可知, 在目标命中率 P 一定的情况下, 区间大小与该区间有效 ID 的个数正相关, 因此在这里我们可以通过区间内有效 ID 样本数与有效 ID 样本总数的比值来确定该区间内采样次数占总采样次数的比值, 即:

$$\frac{\text{区间内有效ID样本个数}}{\text{有效ID样本总数}}=\frac{\text{区间采样次数}}{\text{总采样次数}}$$

通过上式即可确定区间内的采样次数.

在此, 我们对 UNI64 方法的流程进行总结如下:

(1)收集目标网络的有效 ID 样本数据. 由于 OSN 网络结构与用户 ID 无关, 因此可利用爬虫程序按 BFS 搜索的方法收集用户的朋友列表和关注列表, 以快速获得有效 ID 样本集. (2)根据网络的有效 ID 样本集和用户设定的目标命中率对用户 ID 空间进行划分, 得到有效区间. (3)根据有效区间大小分配区间内随机采样次数, 进行采样并得到最终采样结果.

4 实验与分析

我们在新浪微博和Facebook网络上对UNI64方法的效果进行了检验. 首先, 我们从新浪微博和 Facebook 然后设定 20%为目标命中率对区间进行划分; 最后进行 5 组不同次数的采样实验. 作为对比, 我们还分别在 32 位和 64 位 ID 空间内进行了相同次数的随机采样实验, 实验数据及结果见表 4. 根据实验数据分别绘制采样命中率对比图, 见图 3 与图 4. 从图中我们可以看到UNI64方法在新浪微博和Facebook上的采样结果基本都达到了预先设定的 20%的目标命中率, 并且较 32 位空间内的随机采样都有了不同程度的提高, 尤其是在Facebook的采样结果中体现的比较明显. 而在 64 位 ID 空间内随机采样结果为零, 与我们之前的分析一致.

随后我们又对 UNI64 方法采样结果的质量进行了检验. 检验的方法是通过将 UNI64 方法采样结果的分布情况与 32 位随机采样结果的分布情况进行对比, 从而验证 UNI64 方法采样结果的随机性. 我们以 5×10^8 为单位将[0,5×10^9)区间等分为 10 个区间, 然后统计 UNI64 方法采样结果和随机采样的结果中有效 ID 在各区间内出现的频率, 以及各区间内有效 ID 数量占有效 ID 样本总数的百分比. 图 5 和图 6 为分别为两方法在新浪微博和 Facebook 采样结果的分布对比图. 从图中可以看到 UNI64 方法的采样结果与 32 位随机采样结果的分布基本一致, 这说明 UNI64 方法的随机性也是比较符合实际情况的.

表 4 UNI64、32 位随机、64 位随机采样结果对比数据表

	采样次数	UNI64		32 位随机(RD32)		64 位随机(RD64)	
		命中数	百分比(%)	命中数	百分比(%)	命中数	百分比(%)
新浪微博	20000	3870	19.35	2812	14.06	0	0.00
	40000	7784	19.46	5683	14.21	0	0.00
	60000	11611	19.35	8458	14.10	0	0.00
	80000	15551	19.44	11288	14.11	0	0.00
	100000	19491	19.49	14115	14.12	0	0.00
Facebook	20000	4348	21.74	1440	7.20	0	0.00
	40000	8623	21.56	2891	7.23	0	0.00
	60000	12961	21.60	4385	7.31	0	0.00
	80000	17242	21.55	5793	7.24	0	0.00
	100000	21797	21.80	7204	7.20	0	0.00

图 3 新浪微博采样结果对比

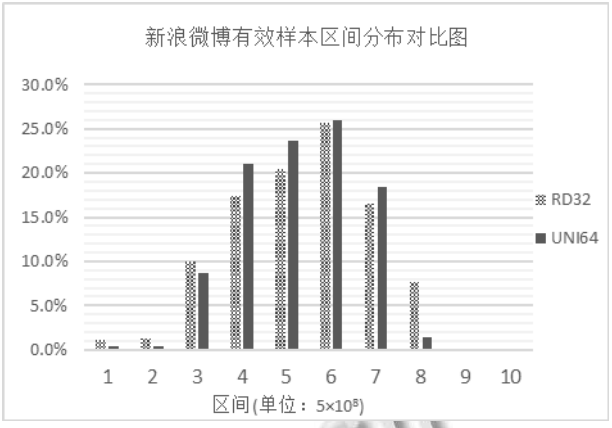

图 5 UNI64 方法与 32 位随机采样(RD32)在新浪微博采样结果分布对比图

图 4 Facebook 采样结果对比

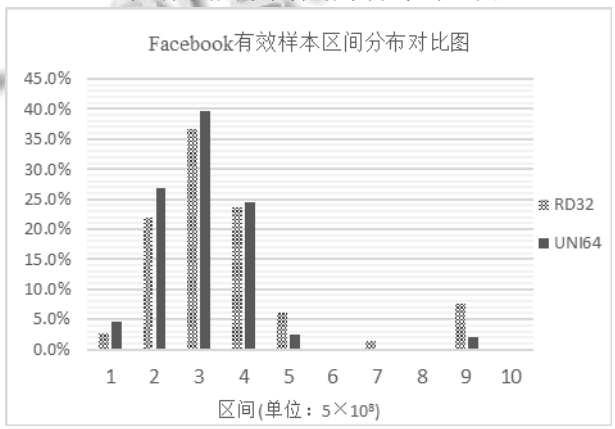

图 6 UNI64 方法与 32 位随机采样在 Facebook 采样结果分布对比图

5 结语

本文为了解决 UNI 采样方法在 64 位用户 ID 系统中命中率急剧下降的问题, 首先对新浪微博的用户 ID 分布进行了分析, 验证了我们对于用户 ID 不均匀分布的假设; 然后提出了一套改进的 UNI 方法, 即 UNI64 方法. 该方法能够通过分析一定量的有效 ID 样本, 在整个用户 ID 系统中划分出能够满足一定目标命中率的有效区间, 并指导 UNI 方法在有效区间内按比例进行随机采样, 从而使实际命中率得到提高, 同时还能保证结果的随机性. UNI64 方法不只适用于 64 位 ID 系统的采样, 还可以推广到其他长度的 ID 系统. 但它也存在两点不足之处: 一是用于划分区间的有效 ID 样本的数量会对最终的采样结果产生影响, 文献[10]中曾指出 15%的网络随机样本已经足够反映网络的基本性质, 但实际应用时, 15%的样本量已经非常庞大. 如果样本数量太少, 又会导致采样结果的分布不足以反映实际 ID 的分布情况. 如何解决这一矛盾, 需要进行进一步研究. 二是目前 UNI64 方法仅采集在线社交网络中单个用户的信息, 而没有获取用户间的关系, 因此样本无法反映实际网络的结构关系. 这也将会是我们在未来工作中的研究方向.

参考文献

1 Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature, 1998, 393: 440–442.

2 Newman MEJ, Watts DJ. Renormalization group analysis of the small-world network model. Physics Letters A, 1999, 263(4–6): 341–346.

3 Barabasi AL, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509–512.

4 Cha M, Haddadi H, Benevenuto F, Gummadi KP. Measuring user influence in twitter: The million follower fallacy. 4th International AAAI Conference on Weblogs and Social Media. 2010.

5 Ahn YY, Han S, Kwak H, Moon S, Jeong H. Analysis of topological characteristics of huge online social networking services. Proc. of the 16th international conference on World Wide Web. 2007. 835–844.

6 Choudhary A, Hendrix W, Lee K, Palsetia D, Liao WK. Social media evolution of the Egyptian revolution. Commun. ACM, 2012, 55(5): 74–80.

7 Steeg GV, Galstyan A. Information-theoretic measures of influence based on content dynamics. Proc. of the 6th ACM International Conference on Web Search and Data Mining. New York, NY, USA. 2013. 3–12.

8 Kurant M, Markopoulou A, Thiran P. On the bias of BFS. International Teletraffic Congress (ITC 22). 2010.

9 Gjoka M, Kurant M, Butts CT, Markoppoulou A. Practical recommendations on crawling online social networks. IEEE Journal on Selected Areas in Communications, 2011, 29(9): 1872–1892.

10 Leskovec J, Faloutsos C. Sampling from large graphs. Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006. 631–636.