

DOI: 10.7652/xjtuxb201502006

面向社交媒体嵌入关系数据感知方法的研究

崔颖安^{1,2}, 李雪³, 夏辉¹, 张德运²

(1. 西安理工大学计算机科学与工程学院, 710048, 西安; 2. 西安交通大学电子与信息工程学院, 710049, 西安; 3. 陕西师范大学国际商学院, 710062, 西安)

摘要: 针对社交媒体数据感知成本高、数据感知效率低等问题,提出了社交媒体嵌入关系多阶段数据感知方法(online social media-multi stage data aware, OSM-MSDA)。该方法以数据感知对象内部关系的分布特征为基础,构造一个具有偏好特征的种子网络;采用 Metropolis-Hastings 方法优先选取数据感知对象中高度节点的邻接关系,快速填充特征网络,实现网络轮廓探测;使用基于马尔可夫生灭机制的延迟拒绝方法控制概率转移核,对局部耦合关系进行修剪,确保连通关系疏密的合理分布。实验结果表明:OSM-MSDA 建立的多阶段渐进数据抽样方法,能够克服已有数据感知方法采集样本的盲目性,在宏观尺度准确、高效的感知社交媒体嵌入关系的社会资本特征,确保特征网络与数据感知对象的结构更具有有一致性,同时还能降低数据的使用成本,将数据处理效率提高 32%~63%。

关键词: 社交媒体;嵌入关系;多阶段;数据感知

中图分类号: TP301 **文献标志码:** A **文章编号:** 0253-987X(2015)02-0031-06

A Research on the Data Aware Method for Social Media with Embedding Relationship

CUI Ying'an^{1,2}, LI Xue³, XIA Hui¹, ZHANG Deyun¹

(1. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China;
2. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
3. Department of International Business, Shaanxi Normal University, Xi'an 710062, China)

Abstract: A multi stage data-aware method for online social media with embedding relationship (online social media-multi stage data aware, OSM-MSDA) is proposed to solve problems of data aware in online social media, such as poor availability, high business cost, and low-efficiency, et al. A seed network with preference characteristics is constructed, and then the Metropolis-Hasting method is used to choose adjacency relation with high degree in data aware population. Finally, the improved Delay-Rejection method is used to regulate the Markov probability transition kernel, and to control the distribution density in local network. Experimental results show that OSM-MSDA gets more precise results for social capital of social media and high-efficiency at macro-level, and overcomes the blindness of existing data aware methods. At the same time, OSM-MSDA ensures the consistency between the characteristics of network and the structure of the data object perception, reduces the cost to use data, and increases the data

processing efficiency by 32%-63%.

Keywords: online social media; embedding relationship; multi-stage; data aware

在组织理论中,社交媒体形成的社会网络被定义为关系嵌入与结构嵌入。关系嵌入是指个体行动者的行为嵌入于个体行动者彼此之间的互动关系网络中,通过社会联结的密度、强度、对称性等要素对个体行动者的决策产生影响。结构嵌入是指由个体行动者涌现而成的凝聚子群进一步镶嵌到更大范围的网络中,依据其在整体网络中所处的位置与其他凝聚子群发生联系,从而促进整体网络的演进与组织^[1-2]。

已有研究表明:社交媒体嵌入关系的形成、演化与应用已成为多个学科共同关注的热点问题^[3-4]。尽管不同学科对于社交媒体的研究主题各不相同,但是这些研究均需使用社交媒体嵌入关系数据作为实证基础。社交媒体嵌入关系除了具有多样、快速、灵活、海量的基本特性以外,还有其自身独有的特点。因此有必要对社交媒体嵌入关系数据感知方法进行专门的研究。

综合国内外相关研究文献,现有数据感知方法的研究主要包括以下3类:

(1)人工数据感知方法。由程序开发人员使用通用的编程语言或特别设计的脚本语言,根据每一个网页的特定结构编写个性化的数据感知包装器。由于包装器的数据感知规则与页面结构具有同一性,数据感知的质量和效率通常都会比较高。该方法的不足是一旦页面发生变化,包装器就失去数据感知能力,需要人工修改,维护成本比较高,不适合大规模商用。

(2)半自动数据感知方法。由于人工构造的数据感知包装器学习成本与维护成本都比较高,半自动数据感知包装器就应运而生,该方法需要一定的人工操作,通过数据标注以辅助包装器的生成。通常这些标注操作都会比较简单,标注员不需掌握程序设计知识即可完成。常用的半自动数据感知包装器分为两类:一类是通过归纳推导构造的包装器,包括模式规则法和模板树匹配法;另一类是通过机器学习法构造的包装器,从网页的特征数据中训练出统计模型,实现数据感知与解析。

(3)全自动数据感知方法。不需要用户参与且不使用人工标注训练样本,就能产生数据感知规则的方法统称为全自动数据感知方法。常用的全自动数据感知包装器分为3类:①基于本体知识的数据

感知包装器;②基于视觉信息的数据感知包装器;③基于重复相似子树识别的数据感知包装器。以上3类方法都能自适应地调整数据感知规则以适应网页结构的变化。

1 多阶段数据感知方法

1.1 总体框架

现有数据感知方法用于社交媒体时,还存在数据感知效率低、成本高、规模难以控制等不足,因此本文提出多阶段社交媒体嵌入关系数据感知方法(online social media-multi stage data aware, OSM-MSDA)。该方法的基本思想是以逐步求精的方式,通过合理控制数据感知的规模,构造一个与数据感知对象具有相同社会资本特征的特征网络。该方法由种子网络初始化、网络轮廓探测、局部关系修剪3部分构成。种子网络初始化使用领域问题研究者提供的初始节点作为种子,以数据感知对象内部的关系为基础,采用加点、加边和重连的方法构造一个既具有偏好连接特征也具有随机性的种子网络。网络轮廓探测使用改进的“滚雪球”方法从数据感知对象中挑选合适的关系,对种子网络进行填充,这些新进入的节点与关系与种子网络中已有的节点组织在一起,形成数据感知对象的整体轮廓。局部关系修剪使用基于生灭链机制的 Delay Rejection 方法对数据感知对象中对应的局部关系进行筛选,调整不同节点之间关系分布的密度,确保特征网络与数据感知对象在细节上更具一致性。

1.2 OSM-MSDA 算法过程

1.2.1 种子网络初始化 本文采用偏好随机网络法构造种子网络,该方法的工作过程如下。

(1)初始化种子。从数据感知对象中等概率的选取 n 个节点, m 条边放入特征网络中构造初始种子网络 G , 令 $p, q, 1-p-q$ 分别代表加点、加边和重连的概率。

(2)加点。从数据感知对象中随机的选择一个新节点添加到特征网络 G 中,若该节点的度是 k_i , 采用式(1)的概率分布规则与数据感知对象中的节点相连, α 为 $[0, 1]$ 之间的任意随机数。

$$p(k_i, \alpha) = \sum_{u \in A_i} k_i + \alpha / \sum_{u \in V} (k_i + \alpha) \quad (1)$$

(3)加边。在数据感知对象中随机选择 l 条边, 这些边按照式(1)的概率分布规则连接到特征网络

对应的关系中。如果特征网络缺少数据感知对象中对应的节点,需要补充相应节点构成对应的关系。

(4)重连。在种子网络中,随机选取任意一个节点 i ,删除该节点所有的关系,而后按照式(1)进行关系重连。特征网络关系的选取仍需以数据感知对象为基准,而后在特征网络内选取与之对应的关系进行关系重构。

使用上述方法构造的种子网络具有以下演化特性:式(2)表示增加一个度为 s 的节点的演化规律,式(3)表示增加 l 条边的演化规律,式(4)表示重连 l 条边的演化规律。

$$\frac{\partial k_i}{\partial t} = s \frac{k_i + \alpha}{\sum_j (k_j + \alpha)} \quad (2)$$

$$\frac{\partial k_i}{\partial t} = l \frac{1}{N} + l \left(1 - \frac{1}{N}\right) \frac{k_i + \alpha}{\sum_j (k_j + \alpha)} \quad (3)$$

$$\frac{\partial k_i}{\partial t} = l \frac{1}{N} + l \left(1 - \frac{1}{N}\right) \frac{k_i + \alpha}{\sum_j (k_j + \alpha)} \quad (4)$$

由上述方程,可得种子网络节点度的演化方程为

$$\begin{aligned} \frac{\partial k_i}{\partial t} &= (l-n) \frac{1}{N} + (l+s+n) \frac{k_i + \alpha}{\sum_j (k_j + \alpha)} - \\ &(l+n) \frac{1}{N} \frac{k_i + \alpha}{\sum_j (k_j + \alpha)} = \frac{l-n}{m_0 + t} + \\ &\frac{l+s+n}{l+3s+\alpha} \frac{k_i + \alpha}{t} - \frac{l+n}{2l+s+\alpha} \frac{k_i + \alpha}{t(m_0 + t)} \approx \\ &\frac{l-n}{t} + \frac{l+s+n}{2l+3s+\alpha} \frac{k_i + \alpha}{t} (t \rightarrow \infty) \end{aligned} \quad (5)$$

对演化方程求解得

$$k_i(t) = A \left(\frac{t}{t_i} \right)^\beta - A + (1-p-q)m \quad (6)$$

$$\begin{aligned} A &= (1-p-q)m + a + \\ &\frac{2((1-p-q)m + l) + a}{(1-p-q)[(1-p-q)m + pl + qn]} \end{aligned} \quad (7)$$

$$\beta = \frac{(1-p-q)m + l + qn}{pl - qn - 2a} \quad (8)$$

若加边、加边与重连在时间 t 内以等概率方式发生,则节点度的概率分布函数为

$$P(k_i(t) < k) = p \left(t_i > \left(\frac{A}{k + A - (1-p-q)m} \right)^{\beta_i} \right) \quad (9)$$

对其求期望可得

$$\begin{aligned} 2 \leq \gamma &= 1 + \frac{3l+3m+n+\alpha}{(1-p-q)m + pl + qn} \leq \\ &3 + \frac{m+n+\alpha}{(1-p-q)m + 2pl + qn} \end{aligned} \quad (10)$$

由式(10)可知节点度的期望 $\gamma \in [2, 3]$,该值表

明种子网络的度分布符合幂律分布,由此可知种子网络内部的关系具有偏好连接特征。

1.2.2 网络轮廓探测 本文采用改进的“滚雪球”方法进行网络轮廓探测,该方法的工作过程如下。

(1)在种子网络中随机选取任一节点作为初始的“雪球”。使用 Metropolis-Hastings 抽样方法^[5]在数据感知对象中抽选与之对应节点的相邻节点填充种子网络。

(2)Metropolis-Hastings 算法需要构造一个具有平稳性的 Markov 链。为了实现这一目标,需要借助分布函数 $q(x)$ 来控制样本点选取。工作过程为:

步骤1 使用提议函数产生新的候选样本;

步骤2 依据式(11)计算样本接受概率,其中 Δ 为二者中较小的值

$$A(X^{(i)}, Y_1) = 1 \Delta \left\{ \frac{\pi(Y_1) q_1(Y_1, X^{(i)})}{\pi(X^{(i)}) q_1(X^{(i)}, Y_1)} \right\} = 1 \Delta \frac{N_1}{D_1} \quad (11)$$

步骤3 以概率 $A(X^{(i)}, Y_1)$ 接受新样本或者以概率 $1-A(X^{(i)}, Y_1)$ 保持原来的样本。

Metropolis-Hastings 方法非常健壮,本文使用种子网络中的度分布构造概率密度函数 $\pi(X^{(i)})$,使用数据感知对象中相邻节点之间的度分布比值构造 $q(X^{(i)}, Y_1)$ 。Metropolis-Hastings 抽样过程如图1所示。

(3)如果入选节点与种子网络中其他节点也存在相邻关系,则在种子网络中补充对应关系,而后继续在数据感知对象中寻找合适的节点。

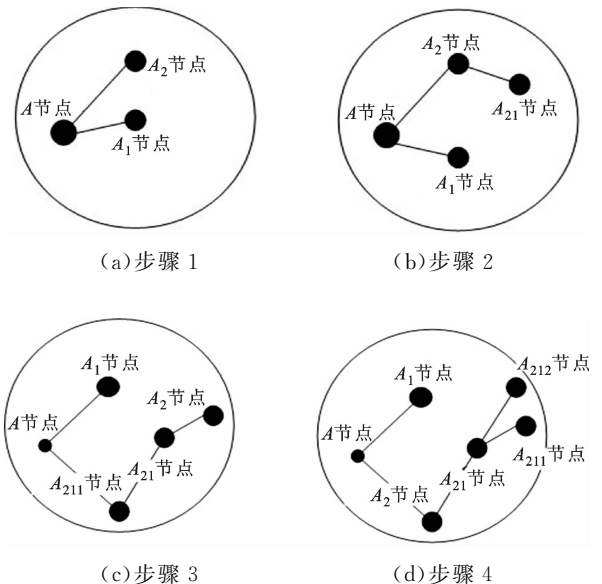


图1 网络轮廓探测中节点关系变化

1.2.3 局部关系修剪 本文使用 Delayed-Rejection(DR)方法进行局部关系修剪,该方法的工作过程如下^[6-7]。

(1)在种子网络中随机选取任一节点作为初始节点。

(2)使用不等概控制方法(如式(13))作为第一层提议函数抽选与之对应节点的相邻节点。如果候选节点选择有效,则将其填充到特征网络中;若无效,将该节点暂时保留到栈中。

$$p(k_i, \alpha) = \frac{\sum_{u \in A_i} 1/k_u}{\sum_{u \in V} (1/k_u)} \quad (13)$$

(3)将舍弃的节点从栈中取出,使用生灭链提议函数作为第二层提议函数,探索该节点的相邻关系,若相邻关系有效,则将其放入特征网络,反之则彻底舍弃该节点。

(4)重复以上过程,直至收敛到平稳分布。

将生灭链方法用于局部关系修剪时,需要知道概率密度函数、提议分布、生灭链状态选择函数以及由此而生的雅克比行列式。通常情况下,获得以上计算要素需进行贝叶斯学习。为了提高效率,本文采用其他方法解决以上问题。

就社交媒体嵌入关系这一特定数据感知对象而言,经过种子网络初始化与网络轮廓探测以后,可以认为特征网络与数据感知对象具有较高的相似性,因此可以使用特征网络的数据近似替代数据感知对象。

另外,MCMC方法有一个非常重要的特性:“提议函数的选取只影响收敛速度,不会影响马尔科夫链最终的收敛”^[8]。因此可以从最“悲观”的情况出发,选择没有稳定期望的柯西分布作为基本提议函数。为了提高提议分布向真实分布逼近的效率,增加样本点与其相继关系的出度比作为柯西分布的修正参数。

对于生灭链转换函数,可以选择正态分布作为生灭链转换函数。正态分布的中心区域代表更新状态,正态分布的两端分别代表死亡和新生的概率(例如生、灭各为7%,更新为86%)。综合基本提议函数、生灭链转换函数以及修正参数共同组成新提议分布。在明确了提议分布以后,雅克比矩阵的计算就非常简单,可由柯西分布和正态分布共同给出。

局部关系修剪中“节点出生”对抽样关系的修正作用如图2所示。在原网络中,图2a所示两个子网之间的有向相干关系分别是2条关系(由左向右)和

3条关系(由右向左),但是经过网络轮廓探测以后,由左向右的连通关系丢失。使用生灭链方法时,假设选中节点 a ,若此时处于“节点出生”状态,则在数据感知对象中,选中 a 的相邻节点 b ,而后根据公式(14)选择其相邻关系(假设 c 被选中),则 $b \rightarrow c$ 关系被选中,将其放入特征网络中以弥补网络轮廓探测的不足。“节点死亡”与“节点出生”相似,是上述过程的逆过程。

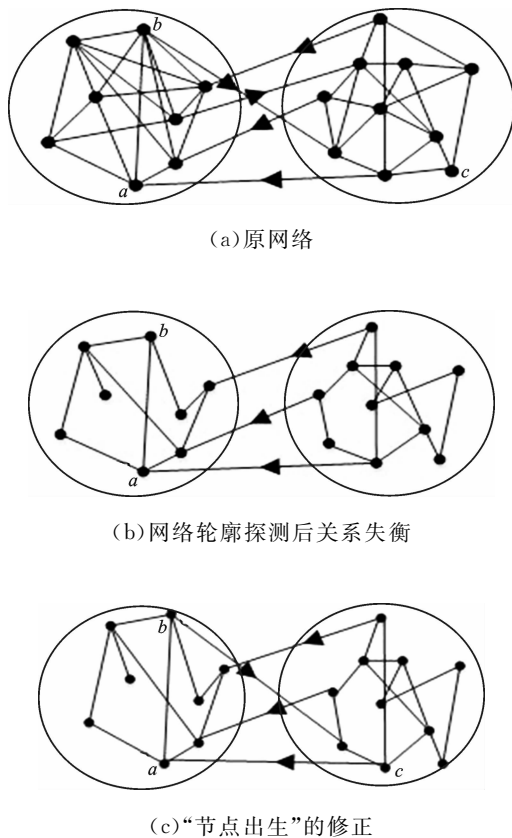


图2 马尔科夫链节点出生变化示意图

2 测试数据集

本文选择新浪微博、蘑菇街、土豆视频、瑞丽作为测试对象。选择以上社交媒体的主要原因是:①系统运营时间长,用户行为趋于稳定,具有研究的稳定性基础;②数据规模庞大,特征网络与数据感知对象对比效果明显;③内部结构复杂,具有测试典型性。测试数据的基本特征如表1所示。

3 数据感知质量分析

3.1 质量特性分析

从整体网络特性、凝聚子群特性、关键节点地位3个维度对特征网络与总体数据集进行比较分析。

表 1 测试数据集

指标	新浪微博	蘑菇街	土豆	瑞丽
节点数	47 058 432	1 124 695	4 391 911	537 590
边数	1 473 399 506	20 249 370	74 662 487	9 147 956
入度系数	-2.432	-1.491	-1.371	-1.293
出度系数	-0.087	-0.169	-0.193	-0.257
相关系数	0.147	0.274	0.089	0.107
平均长度	97.82	41.23	153.67	92.74
簇类系数	0.063	0.127	0.011	0.082

使用 OSM-MSDA 完成数据感知以后,首先需要对特征网络与总体数据的拟合优度进行单样本假设检验。假设检验的真命题是特征网络与数据感知对象的分布具有一致性,假命题是特征网络与数据感知对象的分布不具有一致性。令显著性水平 $\alpha=0.1$,表 2 与表 3 的假设检验结果说明命题真命题是假设成立,表明特征网络与数据感知对象已具有一致性,特征网络数据具有较高的信度和效度。

表 2 OSM-MSDA 宏观特性假设检验

测试对象	入度	出度	相关系数	直径	平均路径长度	簇类系数
新浪微博	0.054 9	0.081 3	0.061 2	0.052 8	0.089 0	0.082 8
蘑菇街	0.096 7	0.077 4	0.090 7	0.087 0	0.106 9	0.099 7
土豆	0.092 1	0.094 8	0.071 9	0.092 5	0.055 2	0.092 2
瑞丽	0.078 5	0.710 4	0.102 7	0.082 9	0.062 1	0.064 9

对采集的数据感知对象总体做进一步的分析,可以发现土豆网的特点是网络规模大,但是其中的孤立点很多(约 35%),缺少明星节点,整体结构类似于随机网络。瑞丽网的特点是网络规模小,有少量的高度节点(约 8%),高度节点的入度虽然较高,但是无标度性不突出,另外低度节点之间的联系很少,整体网络类似若干个小型星型网络组成的复合体。蘑菇街的特点是网络规模大,有一部分高度节点(约 13%),高度节点之间缺少联系,但是低度节

点之间联系比较丰富,整体结构类似于小世界网络组成的复合体。新浪微博的特点是网络规模非常大,内部结构非常复杂,无标度性、小世界性、随机性都比较明显,是典型的混合网络。

表 3 OSM-MSDA 凝聚子群特性 K-S 检验

测试对象	度中心性	接近中心性	介数中心性
新浪微博	0.108 3	0.097 5	0.092 7
蘑菇街	0.092 8	0.097 8	0.104 3
土豆	0.086 2	0.082 1	0.075 1
瑞丽	0.096 5	0.086 5	0.092 3

表 3 给出了特征网络与数据感知对象成分的 K-S 检验结果(如果一个图可以分为多个子图,每个子图内部成员之间有联系,但是不同子图之间没有任何联系,这样的子图被称为成分),表中 K-S 检验结果(除新浪微博 1 个指标以外),均能满足显著性水平检验的要求,表明 OSM-MSDA 对凝聚子群的数据感知效果比较好。另外从表 3 可以看出,内部结构越简单的社交媒体,其假设检验效果越好(令假设检验结果为 I ,则 $I_{土豆网} < I_{瑞丽} < I_{蘑菇街} < I_{新浪微博}$)。此处需要说明的是,此现象会规律性地重复,说明 OSM-MSDA 方法与初始种子的选取无关,方法具有足够的鲁棒性。

综合表 2 与表 3 的数据,可以确定 OSM-MSDA 对不同类型的社交媒体均表现出较好的数据感知效果。从数据感知的运行过程来看:在种子网络初始化阶段,通过有意的控制,确保低度节点与高度节点都能进入特征网络,解决了节点的构成复杂性;在网络轮廓检测阶段,随着高度节点的相邻关系不断进入特征网络,这些样本组织在一起形成多个凝聚子群,解决了节点关系的拓扑复杂性;在局部关系修剪阶段,使用延迟拒绝方法有效的选择低度节点,而后通过马尔科夫生灭链机制调整局部关系分布的密度,形成更大的子群以及整体网络。

表 4 是关键节点的地位关系,表中数据采用皮尔森相关系数对排名前 0.1%高入度与高出度行动者进行了规则相关性分析,根据相关程度将其均分为 4 档。从表 4 中对关键节点地位相似性的 K-S 检验来看,特征网络中高度节点的相似性与真实网络中差异较大,这说明 OSM-MSDA 方法对高度节点关系的数据感知效果存在不足。针对此问题,可以采取扩大样本规模或者使用协方差矩阵构造自适应局部关系分布特征估计函数来优化高度节点的选取。

表 4 关键节点相似性 K-S 检验 %

测试对象	高度相似	中度相似	低度相似	不相似
新浪微博	14.32	21.46	14.73	44.49
蘑菇街	20.11	18.53	24.89	36.47
土豆	17.45	18.42	22.73	43.40
瑞丽	14.55	15.23	16.03	54.19

3.2 性能特性分析

图 3 给出了使用特征网络与总体数据进行社会网络分析耗费时间的对比数据,比较结果显示使用特征网络的分析效率明显优于总体数据,这为领域问题的研究带来了很多方便。事实上,由于社交媒体嵌入关系数据的规模过于庞大(例如新浪微博),使得分析周期过长,就会带来数据分析结果与社交媒体嵌入关系演化不同步的问题,这样的结果对实际工作的参考价值就很有限,甚至有可能出现误导。

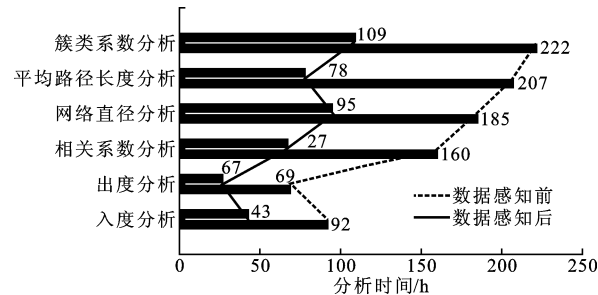


图 3 特征网络与总体数据分析效率对比

4 结 论

本文围绕着总体数据未知,又要快速、低成本地进行社交媒体嵌入关系数据感知这一问题展开了 3 个方面的研究工作。首先对现有数据感知方法进行了分析,指出现有数据感知方法存在的问题;而后提出了多阶段数据感知方法。通过种子网络初始化、网络轮廓探测与局部关系的修剪快速构造了一个与数据感知对象具有较高相似度的特征网络;最后以真实的社交媒体为研究对象,进行了实际测试,测试结果表明 OSM-MSDA 方法具有较好的可用性,能够低成本、高性能为研究者获取社交媒体大数据。

参考文献:

[1] ROOKS G, SNIJDERS C, DUYSTERS G. Ties that tear apart: the social embeddedness of strategic alliance termination [J]. The Social Science Journal, 2013, 50(3): 359-366.

[2] BHARADWAJ A, EL SAWY O A, PAVLOU P A, et al. Digital business strategy: toward a next generation of insights [J]. MIS Quarterly, 2013, 37(2): 471-482.

[3] KITCHIN R. Big data and human geography opportunities, challenges and risks [J]. Dialogues in Human Geography, 2013, 3(3): 262-265.

[4] BESKOS A, CRISAN D, JASRA A. On the stability of sequential Monte Carlo methods in high dimensions [J]. The Annals of Applied Probability, 2014, 24(4): 1396-1445.

[5] MIRA A. On Metropolis-Hastings algorithms with delayed rejection [J]. The American Statistician, 2001, 59(3/4): 231-241.

[6] GREEN P J, MIRA A. Delayed rejection in reversible jump Metropolis-Hastings [J]. Biometrika, 2001, 88(4): 1035-1053.

[7] COTTER S L, ROBERTS G O, STUART A M, et al. MCMC methods for functions: modifying old algorithms to make them faster [J]. Statistical Science, 2013, 28(3): 424-446.

[8] LOVASZ L. Random walks on graphs: a survey [J]. Stochastic processes and their applications, 1974, 2(4): 311-336.

[本刊相关文献链接]

李建东,郑杰,刘勤,等.异构协作网络中采用令牌漏桶的多接入业务分配算法.2014,48(8):7-11.[doi:10.7652/xjtuxb201408002]

安健,桂小林,张进,等.面向物联网移动感知的服务节点发现算法.2011,45(12):6-9.[doi:10.7652/xjtuxb201112002]

杨军,张德运.非均匀分簇的无线传感器网络数据传送机制.2009,43(4):14-17.[doi:10.7652/xjtuxb200904004]

许学斌,张德运,张新曼,等.基于特征层和二代曲波变换的多模生物特征融合识别方法.2009,43(10):32-36.[doi:10.7652/xjtuxb200910007]

王晨旭,秦涛,管晓宏,等.有向网络兴趣社区的快速挖掘算法及其在僵尸粉检测中的应用.2014,48(6):7-12.[doi:10.7652/xjtuxb201406002]

叶娜,赵银亮,边根庆,等.模式无关的社交网络用户识别算法.2013,47(12):19-25.[doi:10.7652/xjtuxb201312004]

张赛,徐恪,李海涛.微博类社交网络中信息传播的测量与分析.2013,47(2):124-130.[doi:10.7652/xjtuxb201302021]

陈国强,王宇平.采用离散粒子群算法的复杂网络重叠社团检测.2013,47(1):107-113.[doi:10.7652/xjtuxb201301021]

(编辑 武红江)