

基于 HMM 的动态社会网络社团发现算法

伊 鹏¹ 周 桥¹ 门浩崧²

¹(国家数字交换系统工程技术研究中心 郑州 450001)

²(中华人民共和国科学技术部信息中心 北京 100000)

(15238363586@139.com)

Dynamic Social Network Community Detection Algorithm Based on Hidden Markov Model

Yi Peng¹, Zhou Qiao¹, and Men Haosong²

¹(National Digital Switching System Engineering & Technology Research & Development Center, Zhengzhou 450001)

²(Department of Information, Ministry of Science & Technology of the People's Republic of China, Beijing 100000)

Abstract With the continuous development of the Internet, most social networks have gradually demonstrated dynamic characteristics, and dynamic analysis of social network community has a very important significance on the understanding of the structure and function of social networks in real life. The HMM_DC algorithm (hidden Markov model based on dynamic community detection) is proposed according to the HMM (hidden Markov model) to detect the community in dynamic social network. Firstly, the algorithm transforms the community detection problem to get the optimal status chain in hidden Markov model considering the history information and characteristics in dynamic social networks. And the algorithm uses the observed chain and status chain to represent the community structure and node information, and can identify the community structure without extra information. Finally, this algorithm and three other algorithms are used to make comparable simulation experiments with VAST data set, ENRON data set and Facebook social network data set. Experimental results show that HMM_DC algorithm performs effectively and accurately in identifying the community structure in the dynamic social network and the value of Q and NMI can be raised greatly compared with other three algorithms.

Key words dynamic social network; hidden Markov model (HMM); optimal status chain; community structure; community detection

摘 要 随着互联网的不断发展,大多数社会网络已逐渐显示出动态特性,动态社会网络社团分析对理解现实生活中社会网络结构和功能具有非常重要的意义.针对动态社会网络中的社团发现问题,提出一种基于隐 Markov 模型(hidden Markov model, HMM)的 HMM_DC 算法.该算法考虑到社会网络的动态特性,结合历史信息,将社团发现转化为求解隐马尔可夫模型中的最优状态序列问题,将网络中的社团结构和节点信息分别采用状态链和观察链表示,在无须指定额外参数的情况下实现动态网络的社团结构发现.最后,利用该算法和其他算法对 VAST 数据集、ENRON 数据集和 Facebook social network

收稿日期:2016-10-11;修回日期:2017-02-06
基金项目:国家“八六三”高技术研究发展计划基金项目(2015AA016102)
This work was supported by the National High Technology Research and Development Program of China (863 Program) (2015AA016102).
通信作者:周桥(zhouqiao_ndsc@163.com)

数据集进行实验仿真. 仿真结果表明: 该算法能够快速、准确地发现真实动态网络中的社团, 其模块度 Q 值和互信息 NMI 值有很大提升.

关键词 动态社会网络; 隐 Markov 模型; 最优状态序列; 社团结构; 社团发现

中图法分类号 TP393

随着互联网技术的快速发展, 各类新型的关于社会网络的应用层出不穷, 并且发挥着日益重要的作用, 社会网络已成为人们日常生活中不可或缺的一部分. 在许多现实和人工的系统中, 社会网络分析是用来研究系统结构和功能特性的主要手段. 近年来, 对众多实际复杂网络的研究发现, 社会网络普遍存在着社团结构这一特性^[1-2]. 同一社团内成员之间具有共同的属性或者在网络中扮演着相似的角色, 联系相对频繁, 社团间成员联系则相对稀疏. 发现和研究网络中的社团结构对于了解网络的内部组织结构和网络功能具有非常重要的意义, 可以分析和挖掘社会网络内部所隐含的模式与信息. 目前很多算法用于发现社会网络中的社团结构, 如基于标签传播算法^[3]、基于模块度优化^[4]的算法以及用于发现网络中重叠社团结构的算法^[5-6]. 值得指出的是, 随着对社会网络研究的深入, 研究者发现在现实生活中, 大多数社会网络都具有动态特性, 网络中的个体随着时间的推移而增加或者减少, 如图 1 所示. 通过对动态网络社团结构特性的研究, 对理解现实生活中社会网络结构和功能具有非常重要的意义.

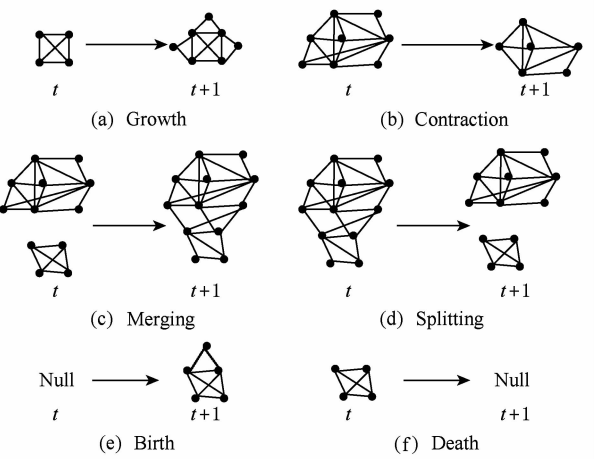


Fig. 1 Community structure in dynamic social network
图 1 动态社会网络中的社团结构

1 相关工作

目前关于发现动态社会网络中的社团结构已经

有了很多成果. 一种主流的对动态社会网络中的结构变化情况进行挖掘的方法是将发生演化的动态社会网络转化为不同时刻的静态社会网络, 如 Berger-Wolf 等人^[7]提出一种分析动态社会网络框架. 该方法着重研究社团结构的演化特性, 主要关注 3 个方面: 1) 社会网络中的结构特性; 2) 发现社会网络中交互关系的变化; 3) 预测社会网络的未来趋势. 动态社会网络中的社团演化情况可以通过社团结构中重大事件的变化来进行分析, 如文献[8]提出的分析动态社会网络中社团结构的 Social Cost 模型, 将动态网络社团发现问题转化为图的着色问题, 但是该方法只关注特定个体的演化关系. 虽然基于上述模型, 动态社会网络中的社团结构发现问题得到了较好的解决, 但仍然存在复杂度较大的问题. 因此, 为了减少复杂度, 提出了其他动态社团发现算法, 如 Sun 等人^[9]提出一种基于信息论的算法 GraphScope, 用来发现社团结构和检测动态网络中发生变化的个体而无需输入任何参数. 文献[10]提出一种用于发现动态有向加权网络社团结构的算法. Nguyen 等人^[11]提出一种自适应增量算法用于发现动态社会网络中的社团结构, 对网络中变化的点和边信息进行处理, 以及文献[12-20]提出了其他相关的动态社团发现方法.

虽然上述方法各有优点, 在某些数据集上表现出较好的计算性能和准确性, 但这些算法只是单独对每个时刻的网络快照进行社团结构发现, 从而找寻各个时刻社团结构之间的演化关系, 忽略了动态社会网络中不同时刻之间的影响关系. 在动态社会网络中, 当前时刻的网络结构信息与前一时刻或者前几个时刻的网络结构信息是相关的, 前一时刻的社会网络中的社团结构对当前时刻社会网络中的社团结构存在着影响, 而之前的动态算法忽略了历史信息对当前时刻网络结构的影响, 使得社团结构的准确性和稳定性偏低.

针对上述动态算法存在的问题, 基于图理论, 从隐 Markov 链模型^[21] (hidden Markov model, HMM) 出发, 本文提出一种基于 HMM 的动态社团发现方法 HMM_DC. 该方法引入 HMM 描述社团结构与

网络中个体间的相互作用,将网络中的社团结构和节点信息分别采用状态链和观察链表示,通过求解 HMM 中的最优状态序列实现动态网络的社团结构发现. 应用于真实网络的实验结果表明,该模型能够在无须指定额外参数的情况下,快速有效地发现动态社会网络内在的社团结构,其模块度 Q 值和互信息 NMI 值得到很大提升.

2 模型建立

如图 2 所示,HMM 是一种描述随机过程统计特性的概率模型,是一种双重随机过程,由 2 个部分组成:Markov 链和一般随机过程. 其中 Markov 链用转移概率来描述状态的转移. 一般随机过程用观察值概率来描述状态与观察序列间的关系.

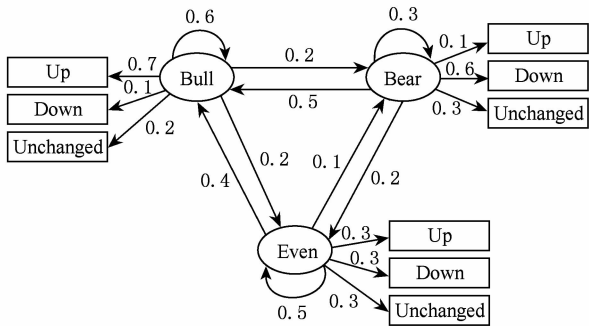


Fig. 2 Example of the HMM model
图 2 HMM 模型示例

本文将动态社会网络按照时间粒度划分为多个网络快照,动态社会网络可由多个时刻的静态网络表示. 假定相邻时刻的网络结构相关,对不同时刻社会网络中社团结构的关系可用一阶 Markov 链描述,由于社会网络中的社团结构不能够被直接观察到,因此,引入隐 Markov 模型将不同时刻社会网络中的社团结构描述为 HMM 中的状态链. 将社会网络中每个节点的节点度分布描述为观察链. 采用隐 Markov 模型描述动态社会网络中的社团结构发现问题,求解隐 Markov 中的最优状态序列得到社团结构,可以更好地理解动态社会网络社团结构演化情况. 相比于当前动态社团结构发现方法,本文在考虑前一时刻社会网络结构信息对当前时刻网络结构信息影响的前提下,将社会网络中社团结构发现问题转化为隐 Markov 模型中的最优状态序列的求解问题. 对于动态社会网络,其社团结构不能够直接被观测出来,而社会网络中各个节点的节点度等属性信息是可以直接观测的. 因此,结合 HMM 模型特

性和上述动态社会网络的特性,提出了一种 HMM_DC 方法用于解决动态社会网络中的社团结构发现问题.

为了得到隐 Markov 模型中的最优状态序列,本文采用 Viterbi 算法^[25]. Viterbi 算法是一种动态规划算法,即给出一个观测序列 $O_1, O_2, O_3, \dots$,找到观测序列背后的隐藏状态 $S_1, S_2, S_3, \dots$,由动态规划寻找出现概率最大的隐藏状态序列.

隐 Markov 模型作为一种具有统计特性的概率模型,从统计意义的角度来描述动态社会网络中的社团发现过程. 现实生活中,社会网络往往有可能存在着噪声,当前动态社团发现方法在噪声的影响下使得结果的准确性降低,但是 Viterbi 算法会查看整个序列来决定最可能的终止状态,然后通过后向指针来找到之前的状态,这对忽略孤立的噪声非常有用,因此该算法对于含有噪声的社会网络同样适用.

3 算法描述

3.1 相关定义

社会网络通常被抽象成图,用 $G=(V, E)$ 来表示社会网络. V 表示社会网络中个体集合, $|V|$ 表示网络中个体的总数; E 表示社会网络中个体间的连接关系集合, $|E|$ 表示社会网络中个体间连接关系集合中元素的总数.

定义 1. 时序图 GS . 时序图 GS_t 表示时刻 t 的社会网络中个体间的关系. 动态网络可以用时序图的集合来表示, $G=(GS_1, GS_2, \dots, GS_t, \dots, GS_T)$, $1 \leq t \leq T$, $\bigcup_{1 \leq t \leq T} GS_t = G$.

如定义 1 所示,大多数研究动态社会网络中社团结构的方法都是将动态社会网络 G 划分成多个时序图 GS 来进行研究. 在结合时刻 t 的时序图 GS_t 仅与前一时刻的时序图 GS_{t-1} 相关这一特性的情况下,建立隐 Markov 模型来发现动态社会网络中的社团结构,给出如下定义.

定义 2. 状态序列 C . 基于动态社会网络 G ,定义动态社会网络中各个时刻的社团结构为隐 Markov 模型中的状态序列 C . 状态序列 $C=(C_1, C_2, \dots, C_t, \dots, C_{N_C})$, 状态数为 N_C , 令 q_t 表示节点在时刻 t 的状态.

定义 3. 可观察序列 O . 基于动态社会网络,定义隐 Markov 模型中的观察序列为网络中每个节点的节点度大小. 观察序列 $O=(d_1, d_2, \dots, d_t, \dots, d_M)$, M 是从每一状态可能输出的不同的观察值数目, d_t 为时刻 t 的某个节点的节点度.

3.2 HMM 模型参数

对于 HMM 模型 $\lambda=(\pi, \mathbf{A}, \mathbf{B})$, $\pi=\{\pi_i\}$ 表示初始状态概率分布, 其中 $\pi_i=P(q_1=C_i), 1 \leq i \leq N_C$ 且 $\sum_{1 \leq i \leq N} \pi_i=1$. $\mathbf{A}$ 为状态转移概率矩阵, $\mathbf{A}=(a_{ij})$, 其中 $a_{ij}=P\{q_{t+1}=C_j|q_t=C_i\}, 1 \leq i, j \leq N_C$, (1) 表示在时刻 t 的状态为 C_i 、时刻 $t+1$ 的状态为 C_j 的概率. $\mathbf{B}$ 为状态 j 的观察概率矩阵, $\mathbf{B}=(b_j(k))$, 表示状态 j 输出相应观察值的概率, 其中: $b_j(k)=P(O_t=d_k|q_t=C_j), 1 \leq j \leq N_C, 1 \leq k \leq M$. (2)

为了得到状态转移概率矩阵 $\mathbf{A}$ 和观察概率矩阵 $\mathbf{B}$, 分别给出如下定义.

设 $\mathbf{x}$ 和 $\mathbf{y}$ 是二元向量, 且 $\mathbf{x}=(x_1, x_2, \cdots, x_n)$, $\mathbf{y}=(y_1, y_2, \cdots, y_n)$, 则 $\mathbf{x}$ 和 $\mathbf{y}$ 向量之间的相似度可定义为

$$Sim(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n x_i \times y_i}{\left(\sum_{i=1}^n x_i^2 \times \sum_{i=1}^n y_i^2\right)^{1/2}}. \tag{3}$$

定义 4. 核心节点. 给出社团结构 $C=(C_1, C_2, \cdots, C_t, \cdots, C_{N_C})$, $\forall C_t, 1 \leq t \leq N_C$. 令 $NCore_{C_t}$ 表示社团 C_t 的核心节点, $Degree(i)$ 为节点 i 的度大小, 则:

$$NCore_{C_t} = \arg \max_{x \in C_t} (Degree(i)). \tag{4}$$

定义 5. 状态转移概率 p_{trans_state} . 对于社会网络中的任意节点 $i \in V, V$ 为节点集合, 给定社团 C_t 的核心节点 $NCore_{C_t}$ 以及节点 i 和核心节点之间的相似度, 则定义矩阵 $\mathbf{A}$ 状态转移概率如式(5)所示. 该状态转移概率表示转移到状态 C_t 的概率.

$$p_{trans_state} = \frac{Sim(i, NCore_{C_t})}{\sum_{k \in (1, N_C)} Sim(i, NCore_{C_k})}, \tag{5}$$

其中, $i \in V$.

在给出观察概率 $p_{observe}$ 之前, 先给出一个隶属函数^[7] $Com(i, C)$ 的定义. 隶属函数的物理意义是: 社会网络的邻接矩阵 $\mathbf{R}$ 中节点 i 隶属于某个群组 C 的程度:

$$Com(i, C) = \frac{1}{Sum\mathbf{R}} \left(r_{ii} + \sum_{j \in C} (r_{ij} + r_{ji}) - \frac{r_{i*}r_{*j}}{Sum\mathbf{R}} \right), \tag{6}$$

其中, r_{ij} 为 $\mathbf{R}$ 中的元素, $Sum\mathbf{R} = \sum_j \sum_i r_{ij}, r_{i*} = \sum_k r_{ik}, r_{*j} = \sum_k r_{kj}$.

定义 6. 观察概率 $p_{observe}$. 给定节点与各个

社团之间的隶属度 $Com(i, C_j), j \in (1, N_C)$, 矩阵中的观察概率为

$$p_{observe} = \frac{Com(i, C_k)}{\sum_{j \in (1, N_C)} Com(i, C_j)}, i \in V. \tag{7}$$

3.3 HMM_DC 算法描述

3.2 节给出了隐 Markov 模型参数的定义, 本节基于隐 Markov 模型介绍发现动态社会网络中的社团结构的 HMM_DC 算法.

HMM_DC 算法的基本思想为: 相比于其他算法, 本算法充分考虑了动态网络中前一时刻网络结构对当前时刻网络结构的影响以及网络中节点和社团结构的固有特点, 建立隐 Markov 模型来解决动态网络的社团结构发现问题. 在当前时刻受到前一时刻网络结构影响的前提下, 基于隐 Markov 链模型描述社团结构与网络个体间的相互关系, 采用状态链和观察链分别表示网络中的社团结构和节点信息, 通过求解隐 Markov 链模型中的最优状态序列得到动态社会网络中的社团结构. 图 3 给出 HMM_DC 算法模型实现框架.

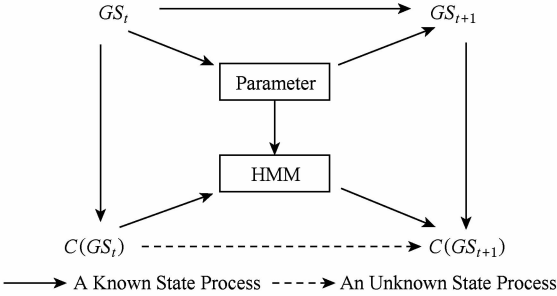


Fig. 3 Algorithm model framework

图 3 算法模型框架

HMM_DC 算法的具体步骤如下:

令时序图 GS_t 中的个体表示为 $(x_1, x_2, \cdots, x_N)$.

步骤 1. 基于静态算法^[4], 对时序图 GS_1 进行社团挖掘, 得到初始社团结构 C_{ini} .

步骤 2. 对于时序图 GS_1 中的节点, 假定其初始状态概率分布为随机分布, 即认为某个节点在初始时刻进入到某个社团的概率是随机的, 得到初始时刻的状态概率分布为 $p_{initial}$.

步骤 3. 在时刻 $t(2 \leq t \leq T)$, 对时序图 GS_t 建立矩阵 $\mathbf{NBC}_{N \times N_C}$, 该矩阵代表节点与社团结构之间的关系, 通过式(4)可以得到每个社团的核心节点 $NCore_{C_t}$.

步骤 4. 在得到每个社团结构的核心节点 $NCore_{C_t}$ 之后, 按照式(3)求得时序图 GS_t 中每个个体 x_i 与各个社团核心节点 $NCore_{C_t}$ 的相似度.

步骤 5. 依照表达式(5)求得时刻 t 的各个状态转移概率 p_trans_state .

步骤 6. 得到时序图中每个个体 x_i 与每个社团结构的隶属度.

步骤 7. 对每个个体 x_i 采用表达式(6)得到社会网络中每个节点隶属于各个社团的隶属度,再采用表达式(7)得到观察概率.

$$NBC_{ij}(t)=\begin{cases}1, & \text{Node } x_i \text{ is in community } C_j. \\ 0, & \text{otherwise.}\end{cases}$$

(8)

步骤 8. 采用 Viterbi 算法来得到最优状态序列,即时刻 t 网络的最优社团结构,令 $t=t+1$.

HMM_DC 算法伪代码如算法 1 所示:

算法 1. HMM_DC 算法.

输入: GS_t 、 GS_{t+1} 、初始社团结构;

输出: 时刻 $t+1$ 最优社团结构 C_{t+1} .

① 初始化:

$len=length(GS_t),$

$NBC=zeros(len,N_c),$

$\Pi=p_initial=Rand(1,N_c);$

② for $1\leq i\leq len$

③ $NBC(i,C(i))=1;$

④ end for

⑤ for $1\leq j\leq N_c$

⑥ 找出每个社团结构的核心节点 $N_{Core};$

⑦ end for

⑧ for $1\leq i_1\leq len$

⑨ for $1\leq j_1\leq N_c$

⑩ 求出相似度 $Sim(i_1,N_{Core});$

⑪ end for

⑫ 计算状态转移概率 p_trans_state 以及矩阵 **A**;

⑬ for $1\leq k_1\leq N_c$

⑭ 计算节点 i_1 与社团结构 C_{k_1} 的隶属度值;

⑮ end for

⑯ 计算观察概率 $p_observe$ 以及观察概率矩阵 **B**;

⑰ 计算 $VITERBI(i,\Pi,p_trans_state,p_observe);$

⑱ if $P=\max(VITERBI)$ (P 为最大概率)

⑲ 将节点 i 加入到使得概率 P 最大的社团结构中;

⑳ end if

㉑ end for

㉒ 得到最终的社团结构 C .

3.4 算法复杂度分析

如算法 1 所示, n 为网络中节点的数目, m 代表网络中边的数目, N_c 为初始社团个数, 行②~④的复杂度为 $O(n)$, 虽然行⑤~⑦代码是针对每个社团进行操作, 但最终仍是对每个节点进行处理, 故复杂度为 $O(n)$; 行⑨~15 的复杂度为 $O(n\times N_c)$; Viterbi 算法采用递归调用, 其算法复杂度为 $O(n^2\times \lg n)$; 故该算法的复杂度为 $O(n)+O(n^2\times N_c)+O(n^2\times \lg n)$. 对于均匀网络, 即在该网络中社团结构不明显, 社团个数 N_c 近似于网络中的节点数, 针对该种网络算法的复杂度为 $O(n^2\times N_c)$; 对于结构比较明显的网络, 由 4.1 节给出的数据集统计得到表 1.

Tabel 1 Data Set Statistics

表 1 数据集统计

Data Set	n	$\lg n$	N_c
ENRON	152	2.1	12
VAST	400	2.6	65
Facebook	60 000	4.7	50

由表 1 可知, 平均社团个数 N_c 要远小于节点数, 且 $\lg n\ll N_c$, 故此种情况下, 算法的复杂度为 $O(n^2\times N_c)$. 综上所述, HMM_DC 算法的时间复杂度为 $O(n^2\times N_c)$.

4 仿真结果及分析

为了验证该模型的有效性, 本文将分别与 CDBIA^[22], QCA^[11], MIEN^[23] 算法进行对比, 仿真实验环境为: 8 GB 内存, 3.2 GHz Intel Core i7 处理器, Matlab R2015b 集成环境, Mysql 5.6 数据库. CDBIA 算法是一种增量社团发现方法, 将动态网络中的变化信息转变为点与边的变化信息进行处理得到不同时刻的社团结构; QCA 算法是一种自适应无监督的用于发现动态网络社团结构的算法, 分别对网中的变化信息进行处理得到各个时刻的社团结构; MIEN 算法将网络中的节点进行压缩和解压缩来描述网络中的变化信息, 并采用快速模块度方法来更新社团结构.

4.1 数据集

为了测试提出的算法在实际数据集上的效果,

采用 VAST 数据集、ENRON 数据集和 Facebook social network 数据集进行实验仿真. VAST 数据集来自 IEEE VAST 2008, 是一个开放的竞赛项目的数据集, 它包括一组共 400 人的 10 d 的通话数据; ENRON 邮件网络数据集是来自 ENRON 公司内部的邮件联系网络, 包含用户邮件消息的网络数据集, 其时间跨度为 1999-05—2002-03; Facebook social network 数据集^[24]是一组 2006-09—2009-01 期间 Facebook 中的好友关系数据集.

4.2 评价指标

为了对社团结构结果进行评价, 采用互信息值 $NMI^{[25]}$ 、模块度 Q 函数^[26] 来进行评价. 其中, NMI 函数用来描述发现的社团结构结果与真实划分结果之间的相似性, 它将互信息的值除以由计算互信息的双方信息量所决定的量; 模块度 Q 函数定量地描述网络中的社团, 衡量网络社团结构的划分, 通过比较网络中各社团的边密度和随机网络中对应子图边密度之间的差异来度量社团结构的显著性.

给定网络中真实社团 U 以及其划分结果 V , 定义 N 为混淆矩阵, 其中行为真实社团, 列为划分所得社团, $N_{i.}$ 为矩阵 N 中第 i 行的和, $N_{.j}$ 为矩阵 N 中第 j 列的和, 真实社团中的个数为 C_U , 划分结果的社团数为 C_V . 则 $NMI(U, V)$ 定义为

$$NMI(U, V) = \frac{-2 \sum_{i=1}^{C_U} \sum_{j=1}^{C_V} N_{ij} \lg \left(\frac{N_{ij} N}{N_{i.} N_{.j}} \right)}{\sum_{i=1}^{C_U} N_{i.} \lg \left(\frac{N_{i.}}{N} \right) + \sum_{j=1}^{C_V} N_{.j} \lg \left(\frac{N_{.j}}{N} \right)}.$$

(9)

在每个时间粒度划分得到的所有社团中, 找到能够被其结构中的任一社团所包含的规模最大的社团. 以此社团为真实社团, 顶点数超过此社团中的顶点都被视为错误划分, 小于此标准的社团则只将不在真实社团中的顶点视为错误划分.

模块度是由 Newman 和 Girvan 提出的一种用于评价划分结果的重要指标^[26], Q 函数定义为

$$Q = \frac{1}{2m} \sum_v \sum_\omega \left[A_{v\omega} - \frac{k_v k_\omega}{2m} \right] \delta(c_v, c_\omega), \quad (10)$$

其中, m 为边的数目; k_v, k_ω 为节点 v, ω 的节点度; $A_{v\omega}$ 为邻接矩阵 A 中的元素; c_v, c_ω 是节点 v, ω 所属的社团.

4.3 仿真结果

1) VAST 数据集. 对该数据集作如下处理: 以每一天为一个时间粒度, 得到 10 个时刻的时序图,

对所得到的时序图采用 HMM_DC 算法, 得到结果如图 4~6 所示:

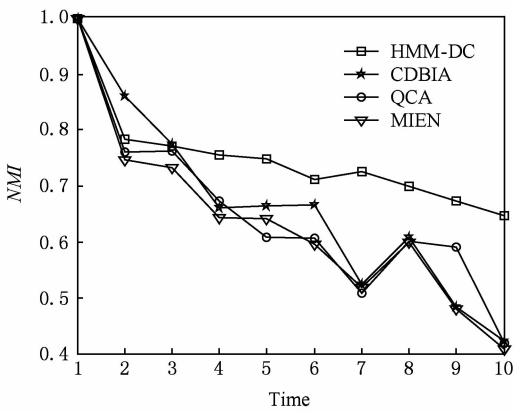


Fig. 4 NMI values of VAST data set
图 4 VAST 数据集的 NMI 值

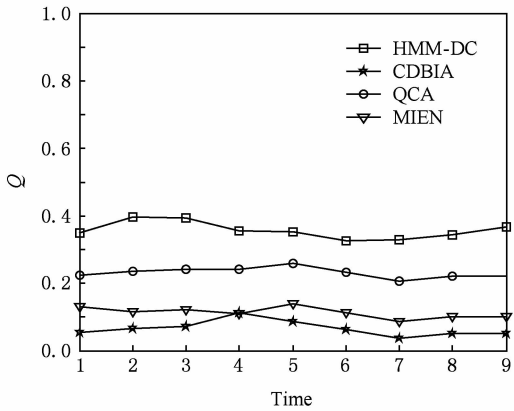


Fig. 5 Q values of VAST data set
图 5 VAST 数据集的 Q 值

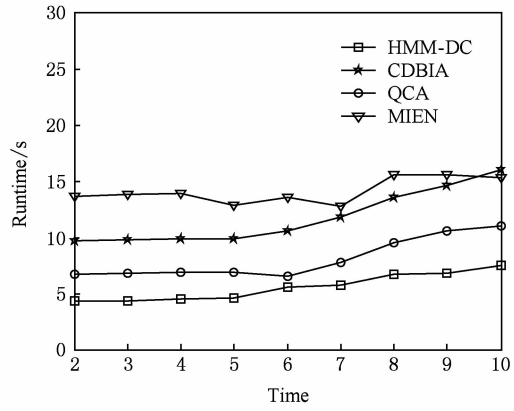


Fig. 6 Runtime of VAST data set
图 6 VAST 数据集的运行时间

图 4~6 分别为 HMM_DC 算法模型在 VAST 数据集中 NMI 值和模块度 Q 值以及运行时间的曲线图. 在图 4 中, 虽然在第 2 个时刻互信息 NMI 值降低了 0.08, 但是在整体上提高了 40%. 在图 5 中,

HMM_DC 算法要明显优于其他算法. 由于 HMM_DC 算法考虑了所有节点在前一时刻的网络结构对当前时刻的影响信息, 而且基于隐 Markov 模型的动态社会网络社团发现方法采用了 Viterbi 算法求得最终的社团结构, 避免了网络中噪声对算法性能的影响, 所以得到的社团结构是全局最优的, 因此 HMM_DC 算法相比于其他算法在 NMI 值和模块度 Q 值上表现都较为良好. 由图 6 可知 HMM_DC 算法的效率要更优.

2) ENRON 数据集. 对该数据集作如下处理: 以每个月为一个时间粒度, 对该数据集进行划分, 分别得到各个时序图, 对每个时刻的时序图采用 HMM_DC 算法, 得到结果如图 7~9 所示. 图 7~9 分别为 HMM_DC 算法模型在 ENRON 邮件网络中的 NMI 值和模块度 Q 值, NMI 值和 Q 值均大于 0.5, 整体提高了 28%. 可以看出, 由于 HMM_DC 算法得到的社团结构是全局最优的, 因此其 NMI 值和 Q 值都较优. 从图 8 中可以看出, 在第 8, 9 时刻本文算法模块度值为 0.53 和 0.535, 略低于 MIEN 算法, 准确性较差. 这是由于这段时间在 ENRON 公司中公司内部员工之间的邮件数减少, 使得时序图变的较为稀疏, 导致结果略有所降低, 然而整体性能相比于其他算法要好很多. 由图 9 可知, 相比于其他 2 个算法, HMM_DC 算法在算法效率上更优.

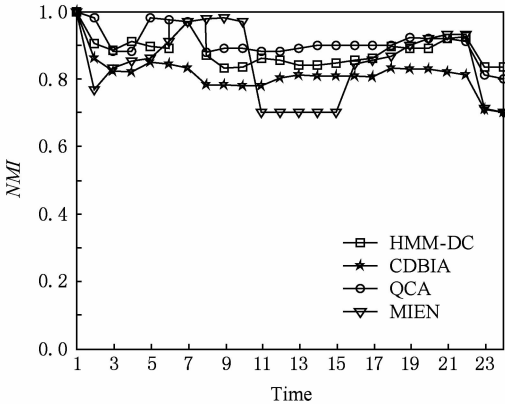


Fig. 7 NMI values of ENRON data set
图 7 ENRON 数据集的 NMI 值

3) Facebook 数据集. 对该数据集作如下处理: 以每一个月为一个时间粒度, 对该数据集进行划分, 分别得到各个时序图, 对每个时刻的时序图采用 HMM_DC 算法, 得到结果如图 10~12 所示. 其中, 图 10~12 分别为 HMM_DC 算法模型在 Facebook 数据集中的 NMI 值和模块度 Q 值以及运行时间的曲线图. 图 10 可以看出, 本文算法在 NMI 值上要

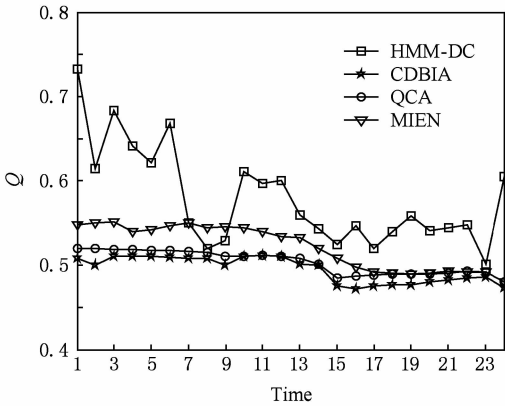


Fig. 8 Q values of ENRON data set
图 8 ENRON 数据集的 Q 值

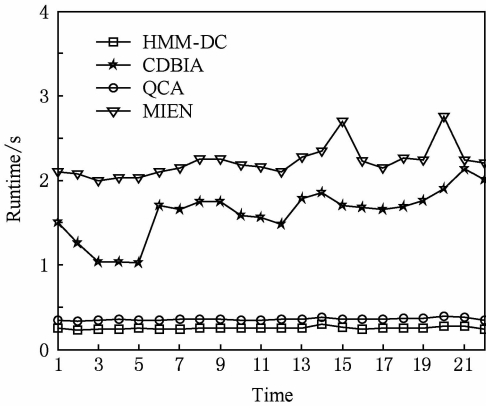


Fig. 9 Runtime of ENRON data set
图 9 ENRON 数据集的运行时间

明显优于其他算法, 平均提高了 20%. 从图 11 可以看出, 模块度 Q 值均大于 0.5, 且相比于 MIEN 和 CDBIA 等其他算法提高了 0.03. 从图 10 和图 11 可以看出, 由于 HMM_DC 算法得到的社团结构是全局最优的, 因此其 NMI 值和 Q 值都较优. 由图 12 可知, 在效率上, HMM_DC 算法与 QCA 算法相似,

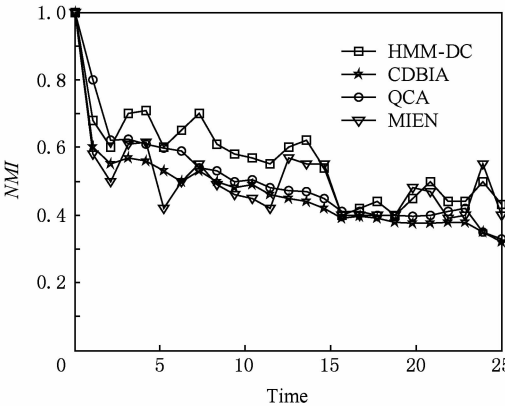


Fig. 10 NMI values of Facebook data set
图 10 Facebook 数据集的 NMI 值

略优于 QCA 算法,与 MIEN 和 CDBIA 算法相比要明显效率高. 因此, HMM_DC 算法在整体性能上要优于其他算法.

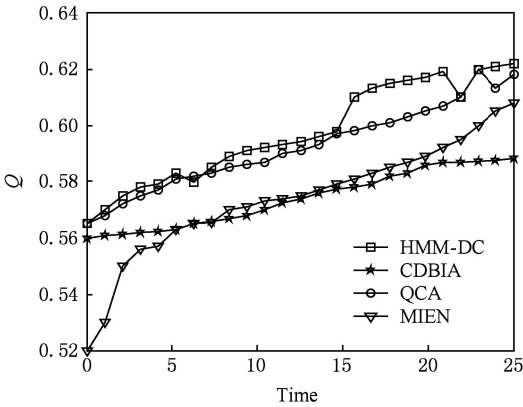


Fig. 11 Q values of Facebook data set
图 11 Facebook 数据集的 Q 值

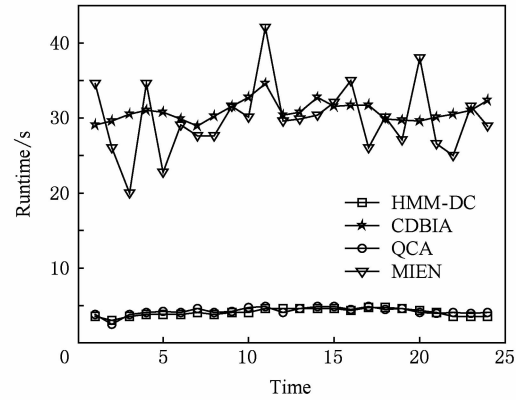


Fig. 12 Runtime of Facebook data set
图 12 Facebook 数据集的运行时间

通过对以上 3 个数据集进行实验仿真可以看出, HMM_DC 算法在准确性和效率上相比于其他动态社团发现方法, 具有较好的性能.

5 结束语

本文提出了一种基于隐 Markov 模型动态社会网络社团发现算法. HMM_DC 通过隐 Markov 模型对动态社会网络进行模型建立, 将动态社会网络中的社团结构建模为 HMM 中的状态链, 而将动态社会网络成员的节点度建模为 HMM 中的观察链, 将求解动态社会网络中的社团结构问题转化为求解隐 Markov 模型中的最优状态序列问题, 最后通过 Viterbi 算法可以得到全局最优的动态社会网络中的社团结构. 通过对真实网络仿真证明了 HMM_DC 算法适用于较大规模且相对稠密的网络, 不仅

能够发现具有实时性、准确性的动态社团结构, 而且具有较低的时间复杂度.

然而, 正如仿真结果中所示, HMM_DC 算法在网络中稀疏图的性能上较差, 由于稀疏图中的节点信息较少, HMM_DC 算法不能得到足够的有效信息而无法得到全局最优的社团结构. 而部分社交网络的动态特性使其在某个时刻具有稀疏性, 因此, 如何改善该算法在稀疏图上的表现是未来将要研究的一个问题.

参 考 文 献

[1] Gregory S. Ordered community structure in networks [J]. Physica A: Statistical Mechanics and Its Applications, 2012, 391(8): 2752-2763

[2] Jin Songchang, Li Aiping, Yang Shuqiang, et al. A MapReduce and information compression based social community structure mining method [C] //Proc of the 16th IEEE Int Conf on Computational Science and Engineering (CSE). Piscataway, NJ: IEEE, 2013: 971-980

[3] Lou Hao, Li Shenghong, Zhao Yuxin. Detecting community structure using label propagation with weighted coherent neighborhood propinquity [J]. Physica A: Statistical Mechanics and Its Applications, 2013, 392(14): 3095-3105

[4] Walker D M, Tordesillas A. Examining overlapping community structures within grain property networks [C] //Proc of 2014 IEEE Int Symp on Circuits and Systems (ISCAS). Piscataway, NJ: IEEE, 2014: 1275-1278

[5] Wu Zhihao, Lin Youfang, Wan Huaiyu, et al. Efficient overlapping community detection in huge real-world networks [J]. Physica A: Statistical Mechanics and Its Applications, 2012, 391(7): 2475-2490

[6] Lin Chuncheng, Liu Wanyu, Deng Derjiunn. A genetic algorithm approach for detecting hierarchical and overlapping community structure in dynamic social networks [C] //Proc of Wireless Communications and Networking Conf (WCNC). Piscataway, NJ: IEEE, 2013: 4469-4474

[7] Berger-Wolf T Y, Saia J. A framework for analysis of dynamic social networks [C] //Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2006: 523-528

[8] Tantipathananandh C, Berger-Wolf T Y, Kempe D. A framework for community identification in dynamic social networks [C] //Proc of the 13th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2007: 717-726

[9] Sun Jimeng, Faloutsos C, Papadimitriou S, et al. GraphScope: Parameter-free mining of large time-evolving graphs [C] //Proc of the 13th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2007: 687-696

- [10] Liu Yao, Liu Qiao, Qin Zhiguang. Community detecting and feature analysis in real directed weighted social networks [J]. Journal of Networks, 2013, 8(6): 1432-1439
- [11] Nguyen N P, Dinh T N, Xuan Y, et al. Adaptive algorithms for detecting community structure in dynamic social networks [C] //Proc of 2011 INFOCOM. Piscataway, NJ: IEEE, 2011: 2282-2290
- [12] Rossetti G, Guidotti R, Pennacchioli D, et al. Interaction prediction in dynamic networks exploiting community discovery [C] //Proc of the 2015 IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining. New York: ACM, 2015: 553-558
- [13] Zhou Xu, Liu Yanheng, Li Bin, et al. Multiobjective biogeography based optimization algorithm with decomposition for community detection in dynamic networks [J]. Physica A: Statistical Mechanics and Its Applications, 2015, 436: 430-442
- [14] Xin Yu, Xie Zhiqiang, Yang Jing. The adaptive dynamic community detection algorithm based on the non-homogeneous random walking [J]. Physica A: Statistical Mechanics and Its Applications, 2016, 450: 241-252
- [15] Liu Ying, Moser J, Aviyente S. Network community structure detection for directional neural networks inferred from multichannel multi-subject EEG data [J]. IEEE Trans on Biomedical Engineering, 2014, 61(7): 1919-1930
- [16] Chen Mingming, Nguyen T, Szymanski B K. On measuring the quality of a network community structure [C] //Proc of 2013 Int Conf on Social Computing (SocialCom). Piscataway, NJ: IEEE, 2013: 122-127
- [17] Cheng Feng, Wang Zuxi. Network community structure in hyperbolic space [J]. Engineering Journal of Wuhan University: Engineering Science, 2013, 46(5): 630-634 (in Chinese)
(程峰, 王祖喜. 双曲空间中的网络社团结构[J]. 武汉大学学报: 工学版, 2013, 46(5): 630-634)
- [18] Shang Jiaying, Liu Lianchen, Li Xin, et al. Targeted revision: A learning-based approach for incremental community detection in dynamic networks [J]. Physica A: Statistical Mechanics and Its Applications, 2016, 443: 70-85
- [19] Wei Wei, Qi Yong. Information potential fields navigation in wireless ad-hoc sensor networks [J]. Sensors, 2011, 11(5): 4794-4807
- [20] Song Houbing, Brandt-Pearce M. Model-centric nonlinear equalizer for coherent long-haul fiber-optic communication systems [C] //Proc of 2013 IEEE Global Communications Conf. Piscataway, NJ: IEEE, 2013: 2394-2399
- [21] Lin Chuncheng, Liu Wanyu, Deng Derjunn. A genetic algorithm approach for detecting hierarchical and overlapping community structure in dynamic social networks [C] //Proc of Wireless Communications and Networking Conf (WCNC). Piscataway, NJ: IEEE, 2013: 4469-4474
- [22] Li Jingyong, Lan Huang, Bai Tian, et al. CDBIA: A dynamic community detection method based on incremental analysis [C] //Proc of IEEE 2012 Int Conf on Systems and Informatics (ICSAI). Piscataway, NJ: IEEE, 2012: 2224-2228
- [23] Dinh T N, Xuan Ying, Thai M T. Towards social-aware routing in dynamic communication networks [C] //Proc of the 28th IEEE Int Conf on Performance Computing and Communications (IPCCC). Piscataway, NJ: IEEE, 2009: 161-168
- [24] Viswanath B, Mislove A, Cha M, et al. On the evolution of user interaction in Facebook [C] //Proc of the 2nd ACM Workshop on Online Social Networks. New York: ACM, 2009: 37-42
- [25] Cui Jinzhong, Zhou Yuanbin, Chen Leiting. Implementation and optimization of embedded speech recognition system based on DHMM [J]. Journal of University of Electronic Science and Technology of China, 2013, 42(6): 930-934 (in Chinese)
(崔金钟, 周远彬, 陈雷霆. 基于 DHMM 的嵌入式语音识别系统的实现与优化[J]. 电子科技大学学报, 2013, 42(6): 930-934)
- [26] Li Xiaojia, Zhang Peng, Di Zengru, et al. Community structure in complex networks [J]. Complex Systems and Complexity Science, 2008, 5(3): 19-42 (in Chinese)
(李晓佳, 张鹏, 狄增如, 等. 复杂网络中的社团结构[J]. 复杂系统与复杂性科学, 2008, 5(3): 19-42)



Yi Peng, born in 1977. PhD, professor. His main research interests include complex network.



Zhou Qiao, born in 1993. Master. His main research interests include computer network, complex network.



Men Haosong, born in 1978. Master. His main research interests include complex network.