



动态社区演化研究进展

潘剑飞, 徐丽丽, 董一鸿

(宁波大学信息科学与工程学院, 浙江 宁波 315211)

摘要:社区结构是社会网络普遍存在的拓扑特性之一。挖掘社会网络中的社区结构、探测并预测社区结构的变化是社会网络研究中重要的研究课题。主要从时间片处理和动态增量的策略对动态社区演化进行阐述, 时间片处理策略介绍了时间片的对比演化、聚类演化、融合演化的研究方法; 动态增量策略描述了核心社区、聚类、指标的动态演化的研究方法; 最后对社区演化预测的框架进行了归纳总结。

关键词:动态社区挖掘; 动态社区演化; 动态社区演化预测

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017026

Research progress of dynamic community evolution

PAN Jianfei, XU Lili, DONG Yihong

Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China

Abstract: Community structure is one of the topological characteristics of social network. It is an important research topic in social network research to explore the structure of community, detect and predict the change of community structure. The evolution of dynamic community was discussed from time slicing processing and dynamic increment strategies. The time slice processing strategy introduced the comparative evolution of the time slice, the clustering evolution of the time slice, and the fusion evolution of the time slice. The dynamic increment strategy described the dynamic evolution of the core community, the dynamic evolution of the cluster and the dynamic evolution of the index. Finally, the framework of community evolution prediction was summarized.

Key words: dynamic community mining, dynamic community evolution, dynamic community evolution prediction

1 引言

社会网络是边与点的连接组成的图, 通过对社会网络的研究, 发现社会网络呈现一种内部顶点连接紧密、外部顶点连接松散^[1]的结构, 即社区结构。社区结构对深入理解和研究社会网络的性质具有很重要的意义, 因此社区研究

将成为真实社会网络分析的一个重要的研究方向。

自 2002 年 Girvan 和 Newman 提出 GN 算法思想^[2]以来, 国内外掀起了社区研究的热潮, 各学科领域 (如生物、物理、计算机等) 的研究者都带来了许多算法思想, 并广泛应用于实际的生活中以解决各个方面的具体问题。随着互联网的快速发展, 更多的人加入互联网中, 新浪、微博和微信

收稿日期: 2016-12-15; 修回日期: 2017-01-06

通信作者: 董一鸿, dongyihong@nbu.edu.cn

基金项目: 浙江省自然科学基金资助项目 (No.LY16F020003)

Foundation Item: Zhejiang Provincial Natural Science Foundation of China (No.LY16F020003)

等交互平台很受广大用户的欢迎,用户交互形成的社交网络存在着显著的社区结构,研究社区结构能够进一步确定用户的社交动态。同时以社区结构的形式发现数据内在的联系,对提高网络管理质量提供辅助决策意见,发现社区内的隐含子群结构特征,发现未来的用户流动、用户需求和商品推荐的研究都有很好的指引作用。

当前对社区结构的研究主要集中针对于静态网络中的挖掘算法设计,随着研究的深入,研究者开始关注复杂动态网络中演化社区结构的识别分析。社区结构识别也正经历从静态研究向动态演化分析的转变,动态网络演化社区结构识别的研究逐步成为研究的热点。

动态社区演化的研究过程可分为时间片处理社区演化和动态增量社区演化。时间片处理社区演化是以时间片为单位,对时间片进行分离、聚类或融合处理,并对其进行时间片社区发现和相似度比较或差距比较两个过程,该做法增加了冗余计算量,同时有些比较策略没有合理的依据。动态增量社区演化解决了时间片可能存在的硬划分问题,通过点边的变化判定前一次社区的变化,能细化到每个社区的演化过程,更具合理性和判定社区动态社区变化的准确性。本文最后引出了社区演化的衍生品社区演化预测的框架,概括地论述了整个动态社区演化研究的主体内容。

2 基本概念

定义 1 (时间片)动态网络 $G_{t:t+n}$ 按时间段分,可表示为 $\{G_t, G_{t+1}, \dots, G_{t+n}\}$, $G_t = (V_t, E_t)$ 为 t 时间段的网络。

定义 2 (时间片社区)对每个时间段的网络 $G_t = (V_t, E_t)$, $t=1, 2, 3, \dots$ 进行社区发现得 $G_t = \{C_{t1}, C_{t2}, \dots, C_{tm}\}$, C_m 表示在 t 时间段的网络中发现的第 n 个社区。

定义 3 (动态网络^[3])在 t 时刻到 $t+n$ 时刻的动态网络 $G_{t:t+n}$ 可表示为 $\{C_t, \Delta C_{t+1}, \dots, \Delta C_{t+n}\}$, $G_t = (V_t, E_t)$ 表示 t 时刻顶点集合为 V_t 、边的集合为 E_t 的网络。 ΔG_{t+1} 表示在 $t+1$ 时刻的更新子图,即 $G_t + \Delta G_{t+1} = G_{t+1}$ 。

定义 4 (动态社区^[3])在 t 时刻到 $t+n$ 时刻的动态社区 $C_{t:t+n}$ 可表示为 $\{C_t, \Delta C_{t+1}, \dots, \Delta C_{t+n}\}$, C_t 表示从 t 时刻网络 G_t 中挖掘的社区。 ΔC_{t+1} 表示在 $t+1$ 时刻的更新子社区,即 $C_t + \Delta C_{t+1} = C_{t+1}$ 。

定义 5 (社区演化事件^[4])对复杂动态网络而言,当网络动态变化时,网络中的社区可能随时间推进而发生持续、缩减、增长、分离、融合、消失、生成等事件,这些事件将导致

社区的结构形态发生动态变化,以下对各个事件进行介绍:

- 持续事件表示随时间推进过程中对应的社区变化很小,社区变化范围一般小于 5%,如图 1(a)所示;
- 缩减事件表示随时间推进过程中对应的社区规模变小,社区变化范围一般为 5%~20%,如图 1(b)所示;
- 增长事件表示随时间推进过程中对应的社区规模变大,社区变化范围一般为 5%~20%,如图 1(c)所示;
- 分离事件表示随时间推进过程中,社区分成多个部分,包括等量分离和不等量分离两种方式,如图 1(d)和图 1(e)所示;
- 消失事件表示随时间推进过程中,对应的社区结构消失,如图 1(f)所示;
- 融合事件表示随时间推进过程中,多个社区发生融合成一个社区,包括等量融合和不等量融合两种方式,如图 1(g)和图 1(h)所示;
- 生成事件表示随时间推进过程中,新的社区结构出现,如图 1(i)所示。

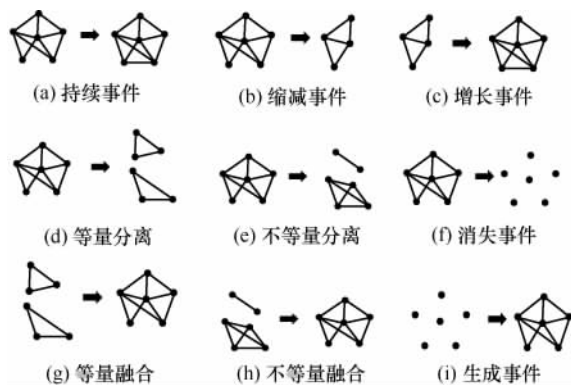


图 1 社区演化事件

3 时间片处理社区演化

时间片处理社区演化是指社区数据按时间段获取形成动态时间片数据,然后以整个时间片为研究单位,对时间片进行分离、聚类或融合等处理后,用静态社区发现算法或静态聚类思想分离出对应的社区,并探索相邻时间片内社区之间的变化,如图 2 所示。

目前时间片分解社区演化的基本策略有 3 种:基于时间片的比对演化、基于时间片的聚类演化、基于时间片的融合演化。

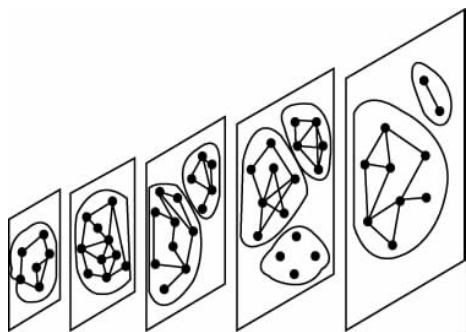


图2 时间片分解社区

3.1 基于时间片的比对演化

基于时间片的比对演化的思想相对比较简单,是基于静态社区发现算法进行的改进,即动态社区发现算法上并没有本质的提高,研究目标从算法研究转为比较策略上的改进。其思想是:首先对整个网络进行时间片处理,然后针对每个时间片,用已存在的静态社区发现算法(LPA 算法^[5]、FCM 算法^[6]、LMF 算法^[7]等)进行社区发现,然后用相似性比较策略对相邻时间片社区进行社区比较,发现社区变化的过程并判定内在的演变事件。

3.1.1 顶点比较策略

参考文献[8]提出在 t 时间片社区 C_t 与 $t+1$ 时间片社区 C_{t+1} 间,顶点变化衡量相似性的比较标准为:

$$\text{sim}(C_t^i, C_{t+1}^k) = \begin{cases} \frac{|V_t^i \cap V_{t+1}^k|}{\max(|V_t^i|, |V_{t+1}^k|)}, & \frac{|V_t^i \cap V_{t+1}^k|}{\max(|V_t^i|, |V_{t+1}^k|)} \geq k \\ 0, & \text{其他} \end{cases} \quad (1)$$

基于此比较标准,结合元社区和二分配的方法得到社区演化序列和演化事件。

- 如果在 $t+1$ 时间片存在社区 C_{t+1}^k 与 t 时间片的社区 C_t^i 匹配,则社区 C_t^i 在 $t+1$ 时间片存活,社区 C_{t+1}^k 将被添加到 C_t^i 对应的元社区中。
- 当 $t+1$ 时间片的社区不断进入,分别对每个社区用最大相似度二分配与 t 时间片的社区进行匹配,如果 $t+1$ 时间片的社区与所有 t 时间片社区都没有匹配,则认为该社区是在 $t+1$ 时间片新形成的社区,对它建立一个新的元社区。

在二分配过程的同时, t 时间片社区 C_t 与 $t+1$ 时间片社区 C_{t+1} 的匹配情况确定演化事件,同时匹配过程中建立的元社区,每个元社区表示为演化的社区序列。

上述比较分析方法的不足在于判定相似度条件上,其

将每个顶点等同看待,这在实际的网络条件下是不合理的,每个社区而言顶点应当有重要程度之分。

3.1.2 核心顶点比较策略

参考文献[9]提出了如何获取核心顶点的算法,核心顶点表示顶点的权值相比其他顶点更高,其中顶点的权值可以由顶点度、PageRank 值等指标来衡量。该算法和投票策略相似,对于每个顶点都有权去评估连接它的核心顶点, $W(N_i)$ 表示顶点 N_i 的度,如果 $W(N_i)$ 的值越大,那么 N_i 点越重要。当 $W(N_i) \geq W(N_j)$, N_i 的核心值 $\text{Cen}(N_i)$ 应该增大 $|W(N_i) - W(N_j)|$ 的值,否则 N_i 的核心值 $\text{Cen}(N_i)$ 应该减少 $|W(N_i) - W(N_j)|$ 的值,通过一轮投票后,当 $\text{Cen}(N_i)$ 的值为非负值时, N_i 为核心顶点,否则为普通顶点。根据以上策略获取到每个时间片的核心点,比较相邻时间片内核心点来确定演化社区序列。

上述比较分析方法的不足在于只考虑社区中的核心点的变化,这种核心点的比较策略没有充分的依据说明方法的有效性,同时没有从整体的角度去考虑社区的演化。

3.1.3 顶点数量与重要程度比较策略

参考文献[4]给出了一种更合理的方法即 GED 算法,该算法根据演化过程中社区顶点数量和顶点重要程度共同的变化衡量相似性。社区顶点的重要程度可以根据 PageRank 的思路进行初始化:

$$NP(x) = (1 - \varepsilon) + \varepsilon \cdot \sum_{i=1}^m NP(y_i \rightarrow x) \quad (2)$$

在 t 时间片中的社区 C_t 中与 $t+1$ 时间片社区 C_{t+1} 的相似度可由 $I(C_t, C_{t+1})$ 和 $I(C_{t+1}, C_t)$ 共同来衡量, $I(C_t, C_{t+1})$ 表示为:

$$I(C_t, C_{t+1}) = \frac{|C_t \cap C_{t+1}|}{|C_t|} \cdot \frac{\sum_{x \in (C_t \cap C_{t+1})} NP_{C_t}(x)}{\sum_{x \in (C_t)} NP_{C_t}(x)} \quad (3)$$

社区演化的事件可以由 $I(C_t, C_{t+1})$ 、 $I(C_{t+1}, C_t)$ 和社区的规模变化共同来决定:

- 当 $I(C_{t+1}, C_t) \geq \alpha$, $I(C_{t+1}, C_t) \geq \beta$, $|C_t| = |C_{t+1}|$ 时,社区 C_t 以社区 C_{t+1} 的形式在 $t+1$ 时间片持续;
- 当 $I(C_{t+1}, C_t) \geq \alpha$, $I(C_{t+1}, C_t) \geq \beta$, $|C_t| > |C_{t+1}|$ 或者 $I(C_{t+1}, C_t) \geq \alpha$, $I(C_{t+1}, C_t) \geq \beta$, $|C_t| \geq |C_{t+1}|$, 并且社区 C_t 在 $t+1$ 时间片只有社区 C_{t+1} 一个社区与之匹配时,社区 C_t 在 $t+1$ 时间片缩减为社区 C_{t+1} ;
- 当 $I(C_t, C_{t+1}) \geq \alpha$, $I(C_t, C_{t+1}) \geq \beta$, $|C_t| < |C_{t+1}|$ 或者 $I(C_t, C_{t+1}) \geq \alpha$, $I(C_t, C_{t+1}) \geq \beta$, $|C_t| < |C_{t+1}|$, 并且社区 C_t 在 $t+1$ 时间片只有社区 C_{t+1} 一个社区与之匹配时,社区 C_t 在 $t+1$ 时间片增长为社区 C_{t+1} ;

- 当 $I(C_t, C_{t+1}) \geq \alpha, I(C_{t+1}, C_t) \geq \beta, |C_t| \geq |C_{t+1}|$ 并且社区 C_t 在 $t+1$ 时间片不止社区 C_{t+1} 一个社区与之匹配时, 社区 C_t 在 $t+1$ 时间片分离;
- 当 $I(C_t, C_{t+1}) \geq \alpha, I(C_{t+1}, C_t) < \beta, |C_t| \leq |C_{t+1}|$ 并且社区 C_{t+1} 在 $t+1$ 时间片不止社区 C_t 一个社区与之匹配时, 社区 C_t 在 $t+1$ 时间片融合;
- 当在 $t+1$ 时间片的每一个社区 $C_{t+1}, I(C_t, C_{t+1}) < 10\%, I(C_{t+1}, C_t) < 10\%$ 时, 社区 C_t 在 $t+1$ 时间片消失;
- 当在 t 时间片的每一个社区 $C_t, I(C_t, C_{t+1}) < 10\%, I(C_{t+1}, C_t) < 10\%$ 时, 社区 C_{t+1} 在 $t+1$ 时间片生成。

GED 算法从顶点数量和顶点重要程度共同来比较两个时间片社区相似度, 在实际运用中的合理性和实用性。

3.2 基于时间片的聚类演化

基于时间片聚类的演化是在静态聚类分析算法的基础上, 提出了一个动态演化聚类思想。参考文献[10]中提出社区在相邻时间片间的变化应该不大, 对此在考察 t 时间片内的社区 C_t 时, 既要满足当前 t 时间片内在的结构, 同时用 $t-1$ 时间片来规整 t 时间片的社区, 并且提出一种衡量 t 时间片社区 C_t 质量的标准, 可以表示为:

$$\text{cost} = \alpha \cdot CS + (1 - \alpha) \cdot CT \quad (4)$$

其中, CS 用于衡量 t 时间片内社区发现是否符合当前社区内部的交互, CT 用于衡量 t 时间片的社区和 $t-1$ 时间片的社区结构的差异, 并用 KL 散度来刻画两者的损失。此方法的思想虽然是时间片处理的策略, 但是在发现社区的结构时提出了用历史社区结构来做参考的思想, 更具合理性, 更符合时间序列的思想。但是, 社区结构的差异和发现过程是不断地对矩阵进行迭代计算过程, 整个过程时间复杂度很高, 并不符合现实生活中真实的大规模网络数据。在参考文献[11]中也提出了此思想, 但是文献属于非重叠社区发现并且局限在规模不可变化的动态社区变化的过程中。

3.3 基于时间片的融合演化

基于时间片融合的演化是将相邻时间片的网络数据进行融合, 并对融合后的数据进行静态社区发现后, 与前一时间片的社区结构进行相似性比较和判定归属分离出当前时刻的社区结构。

参考文献[12,13]中使用到派系过滤算法, 该算法用于重叠最大团社区发现的经典算法(CPM 算法), 通过判断 k 完全图的相互重叠可达性来对最大团进行发现, 在参考文献[14]中提出考虑社区应该考虑其内部的结构分布, 因此,

从内部结构方面来考虑用此方法进行社区发现, 更符合现实的网络数据的结构分布, 并且社区的凝聚程度相对其他策略更高, 但是随着 k 的增大, 发现社区的时间复杂度将急剧增加。时间片融合的重叠最大团社区发现的思想是: 首先将相邻时间的网络 G_t 和 G_{t+1} 合并成为中间网络 G_{t+1} , 然后对 G_{t+1} 网络用 CPM 算法进行社区发现, 结合已存在网络的社区结构, 用顶点的 Jaccard 相似度进行相似度比较, 根据比较结果分离出 G_{t+1} 的社区结构。其顶点的相似度比较如下:

$$S(C, C') = \frac{|C \cap C'|}{|C \cup C'|} \quad (5)$$

其方法在社区发现 CPM 有一定优势的同时, 融合后分离的策略考虑了历史社区的结构分布, 并不是时间片单独处理策略, 解决了时间片孤立处理的硬分割问题。但是假设两个时间片间的社区变化不是很大的情况下, 每次都要进行社区发现, 无形中增加了社区演化发现过程中的时间复杂度。

3.4 时间片策略综合比较

时间片的比对演化、聚类演化、融合演化都属于时间片处理的社区演化策略, 时间片的比对演化思想, 在时间片处理方面过于粗糙, 直接将其时间片分离, 相邻时间片社区划分之间可能呈现显著的结构变化, 属于硬分割问题。时间片的聚类演化思想, 当前时间片社区的划分依据当前时刻社区情况和历史时间片社区的划分来决定, 根据社区在相邻时间片变化不是很大的原理, 该考虑和历史时间片社区的情况的差异最小化的思想具有合理性。时间片的融合演化进行相邻时间片融合然后对重叠社区进行发现虽然解决了硬分割的问题, 但是这 3 个策略不属于真正的动态社区演化, 对每个时间片处理复杂度高, 甚至还得用些相似度比较获取社区演化序列, 然而比较策略标准不一, 没有合理的依据。时间片处理社区演化策略综合比较见表 1, 其中 n 表示网络顶点的数目, e 表示网络边的数目, m 表示社区的数目, d 表示网络的度(顶点度的均值)。

4 动态增量社区演化

动态增量社区演化是指根据初始时刻网络的社区, 随着时间的变化更新社区结构, 从而发现动态社区演化序列和事件。相比于时间片处理的社区演化, 动态增量的社区演化充分利用了前次识别的结果, 在前次的结果上进行动态判定, 真正实现了动态识别和演化。



表 1 时间片处理策略综合比较

策略	时间复杂度	硬分割问题	社区内结构	相似度比较	是否参考历史社区信息
比对演化	一般为 $O(n)$	存在	取决静态社区发现算法	存在	否
聚类演化	$O(e \cdot m)$	不存在	合理	不存在	是
融合演化	$O(d \cdot n^3)$	不存在	合理	存在	是

4.1 基于核心社区的动态演化

基于核心社区的动态演化的思想在于首先判定核心点组成的核心社区的动态变化,然后扩展为整个网络的社区变化。

参考文献 [15] 中提出追踪社区演化的统一框架——FICET(fast increment community evolution tracking,快速增量社区演化跟踪)框架,整个社区演化过程既考虑当前社区结构的合理性,又考虑了当前社区结构和历史社区结构的衔接性;同时,追踪社区演化针对核心社区的增点、增边、减点和减边的 4 种变化的研究,既能保证高效率演化也能快速跟踪社区演化的过程。整个框架如图3 所示。

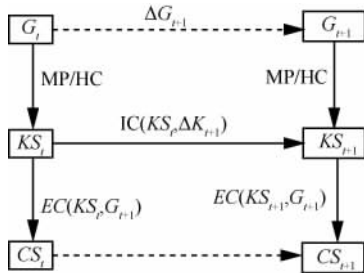


图 3 FICET 框架

整个框架的思想主要分为两个步骤:社区的发现和核心社区动态增量演化扩展。社区发现是动态增量演化的基础,随后的演化都是在此基础上的。社区发现分为 3 个步骤。

步骤 1 通过改变后的 PageRank (modified weighted pagerank, MP)算法,发现网络的核心点、连接边组成核心的子图。利用改进后的 MP 算法求 PageRank 值:

$$PR(v_i) = \left(1 - \frac{1}{w_{v_i} \times \sum_{k=1}^m w_{v_k}}\right) \times \sum_{v_j \in N(v_i)} \frac{PR(v_j)}{w_{v_j}} + \frac{1}{w_{v_i}} \quad (6)$$

步骤 2 对核心的子图,运用高层聚类(high-level clustering, HC)算法发现核心的社区集。其中 HC 算法中引入适应度函数来作为核心社区形成标准:

$$f(C) = \frac{PR_C^{\text{in}}}{PR_C^{\text{in}} + PR_C^{\text{out}}} \quad (7)$$

PR_C^{in} 表示核心社区内部顶点总共的 PageRank 值, PR_C^{out} 表示核心顶点邻边的核心点。判定邻边顶点 v 是否属于核心顶点即考虑 $f(v) = f_{C+[v]} - f_{C-[v]}$ 的值,当 $f(v) > 0$ 时,顶点被添加在核心社区中。

步骤 3 对核心的社区集,运用扩展(core community expanding, EC)算法扩展为整个网络的社区。其中 EC 算法引入顶点和各个核心社区间亲密度的概念,以顶点和各个核心社区间的大小来硬分割社区。顶点和核心社区的亲密度为:

$$f_{\text{in}} = \frac{\sum_{u \in (N(u) \cap C)} PR(u)}{\sum_{u \in C} PR(u) + \sum_{u \in (N(u) \cap C)} PR(u)} \quad (8)$$

社区发现后,将经历核心社区动态增量演化扩展的过程,增量演化运用增量(incremental community evolution, IC)算法对前一时刻的核心社区进行增量变化来发现当前核心社区。同时引入社区结构稳定参数(community structure stability, CSM)来防止数据漂移的情况。扩展过程和社区发现的扩展过程一致。其中,IC 算法考虑核心顶点集和对应的边集的变化:

$$V_{t+1} = V_t - V_{\text{del}} + V_{\text{new}} \quad (9)$$

$$E_{t+1} = E_t - E_{\text{del}} + E_{\text{new}} \quad (10)$$

其中, V_{del} 表示在 $t+1$ 时刻消失的顶点, V_{new} 表示在 $t+1$ 时刻新增加的顶点, E_{del} 表示在 $t+1$ 时刻消失的边, E_{new} 表示在 $t+1$ 时刻新增加的边。针对每一种增点、增边、减点和减边的变化来动态判定社区可能发生的变化。增点、增边可能导致社区的增长和社区的融合事件,减点、减边可能导致缩减和消失事件。

快速增量社区演化跟踪框架体现了动态增量社区演化的本质思想由 t 时刻核心点的社区来指导 $t+1$ 时刻核心社区的发现和核心社区演化,然后对演化后核心社区通过扩展算法来获取整个网络内的社区结构。追踪社区演化针对核心点社区的动态变化,既能保证演化的准确性,也能快速跟踪社区演化的过程。但是,核心社区扩展的过程采用的是顶点强硬社区分配,不能对重叠社区进行发现,同时社区结构稳定参数不够合理。

4.2 基于聚类的动态演化

基于聚类的动态演化思想在于将静态聚类的思想和概念运用在动态更新社区和演化事件跟踪的过程。

参考文献 [16] 中基于静态密度聚类思想提出了 DENGGRAPH 的算法。动态演化的过程,考虑积极的改变和消极的改变两个方面。积极的改变为考虑新的核心顶点、新的边界顶点、新的邻接顶点添加时导致原本社区的变化。

- 新的核心顶点,考察是否与已存在的社区存在连接,如果存在,将其添加到连接社区中,否则形成新的社区。
- 新的边界顶点,考察边界顶点是否与社区中的核心顶点密度可达,如果可达,将其添加在社区中,否则作为孤立点处理。
- 新的邻接顶点,将其添加到邻接的社区,并且更新社区内的强度。

积极的改变可产生社区的生成、增长和融合事件,如图 4 所示。

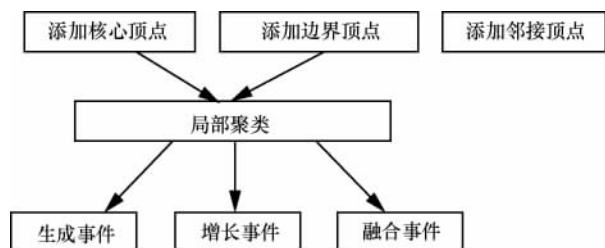


图 4 积极改变

消极的改变为考虑核心顶点、邻边顶点、边界顶点和核心顶点的边,核心顶点间边消失时导致原本社区的变化。

- 核心顶点的消失,考虑核心顶点周围的顶点,判断其与其他核心顶点的可达性来决定原本顶点的归属社区。
- 邻接顶点的消失,当邻边顶点为边界顶点时,更新社区的强度。
- 边界顶点和核心顶点的边消失,原本的边界顶点可能将不可达,边界顶点将成为孤立点。
- 连接核心顶点边的消失,两个核心顶点将不可达,可能产生社区间的分离事件发生消极的改变可产生社区的消失、缩减和分离的事件,如图 5 所示。

采用聚类的思想进行动态社区演化发现,方法简易,能根据静态聚类对应的参数和顶点变化来判定社区的变化和对应的社区事件更具合理性,而且能细化到每个细节的变化可能导致的社区演变事件。

4.3 基于指标的动态演化

基于指标的动态演化的思想主要在于通过点、边社区内和社区间的变化,设定顶点的指标判定顶点归属社区,从而衡量社区的动态演化过程。

参考文献[17]中根据增点、增边、减点、减边和对于顶点是否归属同一个社区,建立候选的顶点集后判定社区演化的强度,如果强度大于阈值时,该时刻社区重新进行社区发现,否则根据候选集中顶点在相邻社区内顶点和邻接

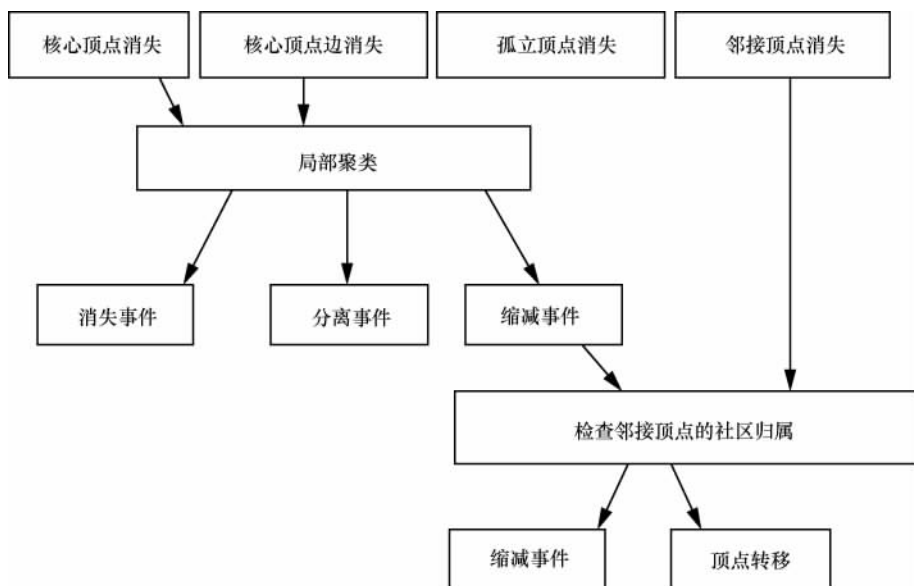


图 5 消极的改变



顶点的永久度大小,动态更新顶点的归属社区,从而动态更新社区。其社区的演化强度可表示为:

$$\text{Strength}(t) = \text{Strength}(t-1) + \frac{(|v_i^+| + |\text{Neighbor}(v_i^+)|) + |\text{Member}(v_i^-)|}{|V_{\max}|} + \frac{2 \times |e_i^+| + |\text{Member}(w_i)|}{|V_{\max}|} \quad (11)$$

社区演化强度本质而言就是不断累加社区点边变化的强度,如果变化强度超过一定范围就认为社区演化可能产生数据漂移现象,应当重新进行社区发现的过程来防止社区演化异常,这是演化事件累积指标,并不能明确指出当前是否已经发现偏移情况。其顶点的永久度可表示为:

$$\text{Perm}(v) = \frac{I(v)}{E_{\max}^{(v)}} \times \frac{1}{D(v)} - (1 - c_{\text{in}}^{(v)}) \quad (12)$$

顶点的永久度考虑两个方面:一个方面是社区顶点的社区内连接和社区外连接比,另一个方面是社区内部结构的结构有效度。

参考文献[17]用顶点的归属社区变化来合理地确定整个动态过程,但是,并没有发现整个社区演化的事件,并且顶点的永久度和社区发现过程也是引用参考文献[18]的思想。

参考文献[19]根据增加、减少离群顶点,增加、减少边,顶点从一个社区转移到另一个社区的5种变化,判断顶点的社区成员度的变化,从而判定顶点的归属社区,进而实现动态社区发现。

4.4 动态增量策略综合比较

动态增量社区演化分为核心社区的动态演化、聚类的动态演化、指标的动态演化3个部分,其区别在于核心社区的动态演化主要通过核心社区和扩展判定社区的演化,在时间复杂度上要优于其他策略,但是从探索社区变化的细节和探索演化事件的准确性方面,聚类和指标的动态变化要高一些,但是依据参考文献[14]中的思想,3种策略欠

考虑社区内部的结构方面,本质以边作为研究对象,把所有的边同等看待,在实际的网络中表现的社区结构并不明显。动态增量社区演化策略综合比较见表2,其中 α 为两个时刻间的变化事件转为顶点的变化比率。

5 社区演化预测

社区演化预测是历史动态社区演化事件判定之后衍生出的新的研究热潮,在已知历史社区本身的特征和历史社区演化事件的基础上,如何探究接下来社区会发生什么样的演化事件,即社区演化预测。社区演化预测的思路具有代表性的是Stanislaw提出的参考文献[20],其他文献在社区演化的思路方面和Stanislaw大体一致,主要还是在时间片处理社区演化的基础上衍生的课题,如参考文献[21]预测演化事件的出发点在于不同的网络,对不同时间片运用不同的静态社区发现算法将导致最终不同的社区演化事件的预测准确性。参考文献[22]预测演化的周期,本质而言就是判定下一时间片社区是否消失,从而确定社区演化周期。Stanislaw提出的社区演化预测思路如图6所示。

社区演化预测思想主要包括以下3个方面:

- 根据以上动态社区演化判定演化序列中两社区间的演化事件;
- 发现社区内部的特征、社区间的特征和演化序列中两个时间段的特征变化;
- 运用合理的机器学习方法进行分类预测。

5.1 特征选取

参考文献[20]中的特征选择主要是从社区整体特征、点特征、点特征量统计计算3个角度,社区整体特征包括社区的大小、社区内点与点之间的连接密度、社区内部的凝聚度、社区核心顶点、社区边的相互作用;点特征包括点的入度、出度,连接顶点的最短路径,点对应的特征向量,顶点与其他社区内顶点距离和;点特征量统计计算包括对所有特征进行求和、最小值、最大值、均值等运算。社区内点与点之间的连接密度为:

表2 动态增量策略综合比较

策略	时间复杂度	适用范围	社区内结构	数据漂移	优缺点
核心社区动态演化	$O(\alpha \cdot n)$	PageRank 思想基础	考虑点的重要程度,未考虑社区内结构	解决	动态演化判定快但不精确
聚类动态演化	$O(\alpha \cdot n \lg n)$	静态密度聚类基础	边为基本单位,社区内结构不能很好地保证	未解决	动态判定细致但把所有边同等看待,社区内结构欠考虑
指标动态演化	$O(\alpha \cdot n \cdot d)$	适用所有社区发现基础		解决	

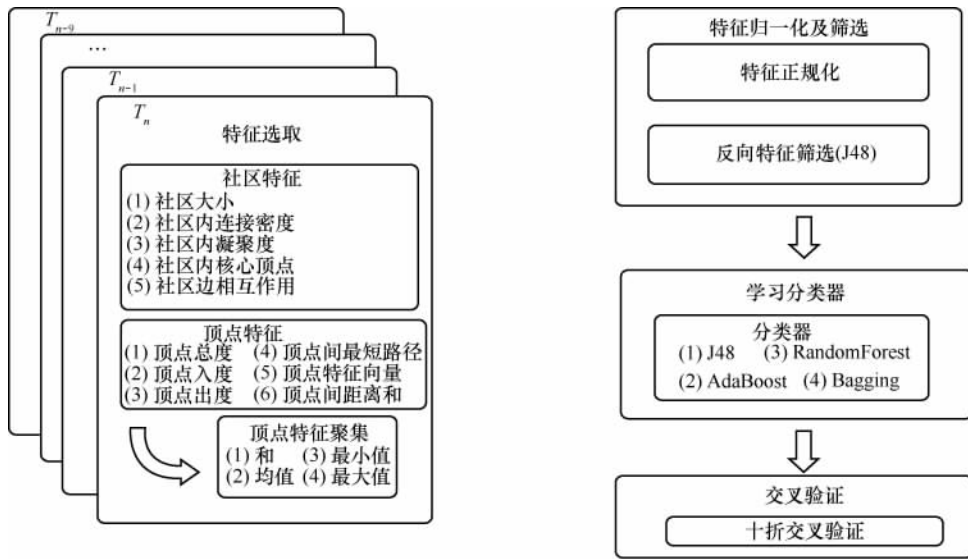


图6 社区演化预测思路

$$D = \frac{\sum_i \sum_j a(i, j)}{n(n-1)} \quad (13)$$

当顶点与顶点之间存在连接时, $a(i, j)$ 为 1。社区内部的凝聚力为:

$$C = \frac{\sum_{i \in G} \sum_{j \in G} w(i, j)}{n(n-1)} \quad (14)$$

其中, $w(i, j)$ 表示顶点 i 与顶点 j 组成边的权值, n 表示在社区内的顶点数, N 表示整个网络内的顶点数。社区边的相互作用为:

$$R = \frac{1}{m} \sum_{ij} a(i, j) a(j, i) \quad (15)$$

其中, m 表示整个网络中的边的数目, 当顶点 i 与顶点 j 之间存在连接时, $a(i, j)$ 为 1。

参考文献[21]中的特征选择包括社区内部的特征、社区间的特征和演化序列中两个时间段的特征变化, 社区内部特征包括核心点的比例、核心点的度、核心点的凝聚度、核心点的特征向量; 社区间的特征包括社区的大小、社区的凝聚度、社区聚类有效性、社区内连接度、社区的度的均值、社区特征向量的均值; 演化过程中两个时间段的特征变化为以社区角度来考察对应社区特征的变化。其社区聚类有效性为:

$$CC(V) = \frac{\# \Delta}{\# Triples} \quad (16)$$

其中, $\# \Delta$ 表示在社区内组成的三角形的数目, $\# Triples$ 表示整个网络中三角形的数目。社区内连接度为:

$$C(u) = \sum_{v \in u} sh(u, v) \quad (17)$$

其中, $sh(u, v)$ 表示社区内两点间组成的最短路径。

5.2 分类预测

参考文献[20]中获取社区的特征后首先进行特征归一化及筛选, 然后运用了机器学习的分类算法 J48、RandomForest、AdaBoost、Bagging、Logistic 进行训练, 对数据集进行十折交叉验证, 得到预测模型, 然后用模型对对应时间片的社区进行预测。参考文献[22]中提出了对每种事件分开进行训练, 分别对每种事件进行判定预测准确性, 同时调用了 weka 的 api 对特征进行优化选取和降维, 分析社区特征对最终预测的重要程度。

由于社区网络的复杂性、基于时间片硬分割分离的社区发现和相似度定义社区历史演化事件的不合理性, 发现社区事件会出现严重的偏移情况, 社区消失的事件比其他社区事件要多很多, 参考文献[21]中提出了针对每个事件的进行 weka 中的 SMOTE 采样, 使训练样本每个事件的训练样本数目达到均值, 然后对每个机器学习分类算法进行训练, 这种策略没有合理的依据, 同时对于以上预测方法, 都是将每个时间片的社区特征归结在一起进行训练, 没有对每个时间片社区演化序列作为训练的对象, 无形中缺少了社区演化序列间的时间特征。



6 结束语

介绍了动态社区演化的研究现状和社区演化预测的相关工作。动态社区演化主要对当今现存的主流的时间片处理策略和动态增量演化策略进行阐述,但是现今研究动态社区演化还处在单机处理阶段。随着大数据时代的来临,更多的研究人员将对社区的动态演化过程进行分布式处理,提高整个演化计算的效率。同时,动态社区演化衍生出的社区演化预测还处于刚起步阶段,没有成熟的理论架构基础,如何建立有效预测模型和统一演化预测过程将成为未来的研究趋势。

参考文献:

- [1] WANG L, CHENG X Q. Dynamic community online social network discovery and evolution [J]. Chinese Journal of Computers, 2015, 38(2): 219-237.
- [2] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of sciences of United States of America, 2002, 99(12):7821-7826.
- [3] PEI L, LAKSHMANAN L V S, MILIOS E E. Incremental cluster evolution tracking from highly dynamic network data [C]// IEEE International Conference on Data Engineering, March 31-April 4, 2014, Chicago, USA. New Jersey: IEEE Press, 2014: 3-14.
- [4] BRÓDKA P, SAGANOWSKI S, KAZIENKO P. GED: the method for group evolution discovery in social networks [J]. Social Network Analysis and Mining, 2013, 3(1): 1-14.
- [5] WU T, CHEN L, GUAN Y, et al. LPA based hierarchical community detection [C]//IEEE International Conference on Computational Science and Engineering, August 22-24, 2014, Vancouver, Canada. New Jersey: IEEE Press, 2014: 185-191.
- [6] ZHANG S, WANG R, ZHANG X. Identification of overlapping community structure in complex networks using fuzzy C-means clustering [J]. Physics A: Statistical Mechanics and its Applications, 2007, 374(1): 483-490.
- [7] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2009, 11(3): 19-44.
- [8] TAKAFFOLI M, SANGI F, FAGNAN J. Community evolution mining in dynamic social networks [J]. Procedia-Social and Behavioral Sciences, 2011, 22(22): 49-58.
- [9] WANG Y, WU B, DU N. Community evolution of social network: feature, algorithm and model[J]. Physics, 2008(3): 31-46.
- [10] LIN Y R, CHI Y, ZHU S. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks [C]//17th International Conference on World Wide Web, April 21-25, 2008, Beijing, China. New York: ACM Press, 2008: 685-694.
- [11] CHAKRABARTI D, KUMAR R, TOMKINS A. Evolutionary clustering [C]//12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 5-8, 2011, Philadelphia, PA, USA. New Jersey: IEEE Press, 2011: 332-337.
- [12] PALLA G, VICSEK T. Quantifying social group evolution [J]. Nature, 2007, 446(7136): 664-7.
- [13] PALLA G, BARABÁSI A, VICSEK T. Community dynamics in social networks [C]//SPIE 4th International Symposium on Fluctuations and Noise, International Society for Optics and Photonics, July 2-August 10, 2007, San Diego, USA. New Jersey: IEEE Press, 2007: 273-287.
- [14] PRAT P, REZ A, DOMINGUEZ-SAL D, et al. Put three and three together: triangle-driven community detection [J]. ACM Transactions on Knowledge Discovery from Data, 2016, 10(3): 1-42.
- [15] LIU Y, GAO H, KANG X. Fast community discovery and its evolution tracking in time-evolving social networks [C]//IEEE International Conference on Data Mining Workshop, November 14-17, 2015, Atlantic, NJ, USA. New Jersey: IEEE Press, 2015: 13-20.
- [16] FALKOWSKI T, BARTH A. Studying community dynamics with an incremental graph mining algorithm [C]//Learning from the Past & Charting the Future of the Discipline Americas Conference on Information Systems, August 14-17, 2008, Toronto, Canada. New Jersey: IEEE Press, 2008.
- [17] LI X, WU B, GUO Q, et al. Dynamic community detection algorithm based on incremental identification [C]//IEEE International Conference on Data Mining Workshop, March 3-11, 2015, Toronto, Canada, New Jersey: IEEE Press, 2015: 900-907.
- [18] CHAKRABORTY T, SRINIVASAN S, GANGULY N. On the permanence of vertices in network communities [C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2014, New York, NY, USA. New York: ACM Press, 2014: 1396-1405.
- [19] LI J, HUANG L, BAI T. CDBIA: A dynamic community detection method based on incremental analysis [C]//International Conference on Systems and Informatics, May 19-21, 2012, Yantai, China. New Jersey: IEEE Press, 2012: 2224-2228.
- [20] SAGANOWSKI, STANISLAW. Predicting community evolution in social networks[J]. Entropy, 2015, 17(5): 3053-3096.
- [21] SHAHRI M, GUNASHEKAR S, DOMARVS M V, et al. Predictive analysis of temporal and overlapping community structures in social media [C]//International Conference Companion on World Wide Web, April 11-15, 2016, Montreal, Canada. New Jersey: IEEE Press, 2016.

- [22] TAKAFFOLI M, RABBANY R, ZAIANE O R. Community evolution prediction in dynamic social networks[C]//IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, August 17–20, 2014, Beijing, China. New Jersey: IEEE Press, 2014: 9–16.

[作者简介]



潘剑飞(1991–),男,宁波大学信息科学与工程学院硕士生,主要研究方向为大数据、数据挖掘。



徐丽丽(1993–),女,宁波大学信息科学与工程学院硕士生,主要研究方向为大数据、数据挖掘。



董一鸿(1969–),男,博士,宁波大学信息科学与工程学院教授,主要研究方向为大数据、数据挖掘和人工智能。