doi: 10.3969/j.issn.1005-3026.2015.11.006

一种基于图压缩的重叠社区发现算法

赵宇海, 印 莹, 王 雪
(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘 要: 为提高单机处理复杂网络规模的能力,提出一种新的重叠社区发现算法. 首先,通过基于图压缩的社区结构表示模型(压缩社区图),对网络进行无损压缩;然后,在压缩社区图上基于种子迭代的思想,通过不断优化社区适应度函数将种子扩展成社区;最后,将相似度高的社区进行合并,得到最终的重叠社区结果. 由于压缩后的凝聚图大大降低了待处理的网络规模,并能在一定程度上减少重复计算,该方法可以大大提高计算效率和单机处理的网络规模.

关 键 词: 重叠社区;社会网络;数据挖掘;聚类;图压缩
中图分类号: TP 311 **文献标志码**: A **文章编号**: 1005-3026(2015)11-1543-05

A Graph Compression Based Overlapping Communities Detection Algorithm

ZHAO Yu-hai, *YIN Ying*, *WANG Xue*
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: ZHAO Yu-hai, E-mail: zhaoyuhai@ise.neu.edu.cn)

Abstract: To improve the capacity of single machine to handle complex network, overlapping communities detection algorithm was proposed. First, a graph compression based social network model, namely agglomerative graph, was introduced, which was a lossless compression to the original network. Then, inspired by the idea of iteration based on seeds, the selected seeds were expanded to the communities by optimizing the proposed community fitness function iteratively. Finally, the communities of high similarity with each other were merged to get the final results. Since the scale of the network to be dealt is significantly reduced, and some redundant computations are avoided, the proposed algorithm is of high efficiency.

Key words: overlapping community; social network; data mining; clustering; graph compression

近年来,社区发现不仅成为计算机领域中最具挑战性的基础性研究课题之一,同时也吸引了来自数学、生物、社会学和复杂性科学等其他众多领域的研究者,掀起了一股研究热潮. 很多现实社会网络中,不同社区间往往不是相互独立的,而是彼此重叠、相互关联,每个成员可以同时属于多个社区. 一般称这样的社区结构为“重叠社区”. 重叠社区是对网络的一种覆盖,它反映了更加真实的网络结构,发现重叠社区对研究真实网络的拓扑结构具有更重要的指导意义[1-6].

Palla 等于2005年提出了首个重叠社区发现算法,即基于 k-clique 派系过滤的 CPM 算法[3]. Shen 等[7]提出的基于合并相似极大派系的层次重叠社区发现算法;Lee 等[8]提出的通过对派系进行贪心扩展来获得重叠社区结构的算法. 除了“基于派系”的方法外,Evans 等[9]提出了基于边图转换的方法. 即将原始图转换成边图(line graph),然后在边图上对结点进行聚类,再将聚类结果还原到原始图上. 这种方法在对结点进行聚类时,结点间的相似度是在边图上计算的,这样会因为过度依赖所对应原始图中的两条边的公共结点的度数而导致计算出的相似度值偏高. Lim

收稿日期: 2015-03-30
基金项目: 国家自然科学基金资助项目(61272182);中央高校基本科研业务费专项资金资助项目(N130504001);国家自然科学基金重点资助项目(61332014).
作者简介: 赵宇海(1975-),男,辽宁辽阳人,东北大学副教授.

等[3]对该方法在聚类上进行了改进,提出了LinkSCAN算法.综上,虽然在一定程度上可以发现社区,然而,算法存在参数敏感或准确度不高等情况.另一方面,在网络中寻找派系是费时的过程,因此目前的主流重叠社区发现算法在面对大规模稠密网络时,计算复杂度太高,效率偏低.本文主要研究这样一种算法,在针对大规模复杂网络时,在保证准确度的情况下,提高单机处理的网络规模.

本文设计了一种基于压缩社区模型(压缩社区图)CCG;提出了一个可以在压缩社区上直接进行重叠社区发现的算法 CLEAR;算法在真实数据集和人工数据集的比较分析,证明了提出算法的有效性和高效性.

1 CCG 图

与传统图一样,压缩社区图(compression community gragh,CCG)也有结点和边,但分别是压缩结点和压缩边.简单而言,可以通过对原始图中的结点进行聚类来得到压缩结点,一个类即是一个压缩结点,然后再通过给压缩结点之间增加压缩边,最终得到能够表示原始图的压缩图.压缩图可以大大降低网络的复杂程度,而且能更直观地观察到网络中存在的一些特殊结构.

1.1 基本概念和性质

定义 1 给定图 $G=(V,E)$,其中 V 是结点的集合;$E\subseteq V\times V$ 是边的集合.图 G 对应的压缩图表示为 $G'=(V',E')$,其中 $V'=\{v'_1,\cdots,v'_n\}$ 是压缩结点的集合,满足 $\cup_{v'\in V'}v'=V$;$E'\subseteq V'\times V'$ 是压缩边的集合.并且,如果两个压缩结点 $v'_i,v'_j\in V'$(其中 $i\in[1,n]$, $j\in[1,n]$)之间有一条压缩边 $e'\in E'$ 相连,则这两个压缩结点要么 $v'_i\cap v'_j=\varnothing$,要么 $v'_i=v'_j$.

根据定义 1 得到的压缩图具有 3 种基本结构:"二分派系"结构、"星形"结构和"派系"结构,本节分别加以详细介绍.

1)"二分派系"结构.该结构中,两个压缩结点由一条压缩边连接,对应原始图中一个压缩结点中的所有结点与另一个压缩结点中的所有结点均两两相连,但各自内部的所有结点互不相连.

2)"星形"结构.该结构是"二分派系"结构的一种特殊情况,即其中的一个压缩结点是一个单点集.网络中存在"星形"结构,通常意味着该单点集极有可能是连接若干个不同区域的枢纽.

3)"派系"结构.如果一个压缩结点与自身通过一条压缩边相连,该结构便是"派系"结构.对应到原始图,该压缩结点内的所有结点两两互连,即在原始图中这些结点构成了一个完全子图.

压缩图的生成分为两个阶段:生成压缩结点和压缩边.以下分别加以描述.

1.2 生成压缩结点

本文根据结点的邻居相似度来选择哪些结点可以构成同一个压缩结点.具体来说,给定原始图中的一组结点,如果这些结点有着共同的邻居结点,则它们构成一个压缩结点.基于该度量函数,采用层次聚类,即可生成最终的压缩结点.

定义 2 给定无权图 $G=(V,E)$,结点 u, $v\in V$ 的邻居结点集合分别记为 $N(u)$ 和 $N(v)$.结点 u,v 的邻居相似度 $N_s(u,v)$ 可通过公式(1)来计算:

$$N_s(u,v)=\frac{|N(u)\cap N(v)|}{|N(u)\cup N(v)|}. \tag{1}$$

定义 3 给定带权图 $G=(V,E,W)$,W 为权重集合,每条边 $e\in E$ 被赋予权值 $w\in W$.此时,结点 $u,v\in V$ 的邻居相似度 $N_s(u,v)$ 由公式(2)来计算:

$$N_s(u,v)=\frac{\sum_{x\in N(u)\cap N(v)}\frac{(w(x,u)+w(x,v))}{2}}{\sum_{x\in N(u)\cup N(v)}\frac{(w(x,u)+w(x,v))}{2}}. \tag{2}$$

$N_s(u,v)$ 的取值范围在[0,1]之间.当 $N_s(u,v)$ 为 0 时,说明结点 u 和 v 没有公共邻居结点;当 $N_s(u,v)$ 为 1 时,说明结点 u 和 v 的邻居结点完全相同.

1.3 生成压缩边

上述生成压缩结点的过程可能会导致结果集 C 中存在单点集和较小压缩结点被较大压缩结点包含的情况.若记 U_i 和 W_i 为层次聚类中第 i 层的两个压缩结点,则生成压缩边的过程描述如下.

首先,对压缩结点自身填加压缩边.此过程针对"派系"结构进行.添边过程从最外层压缩结点开始,层层递进由外向内贪婪搜索压缩边.大致过程为:首先判断最外层 U_0 在原图中的导出子图是否是一个完全子图,如果是,则对 U_0 自身添一条边;否则,判断 U_0 的下一层 U_1 在原图中的导出子图是否为完全子图,是则添边,否则继续向里一层判断,重复此过程,直至判断到 U_0 最里层.

其次,为不同的压缩结点添加压缩边.此过程针对"二分派系"结构和"星形"结构进行.大致过

程为：先判断 U_0 与 W_0 所构成的图是否是原图的一个子图，如果是，则在 U_0 与 W_0 之间添边. 当最外层的 U_0 和 W_0 之间有边时，就停止向内搜索，因为这条压缩边就代表了 U_0 和 W_0 内所有结点之间所有的边. 如果最外层的 U_0 和 W_0 之间没有边，再向里一层搜索，先看 U_0 与 W_1 之间是否有边，如果有就添边，并且不再向 W_1 的内层继续搜索. 当 U_0 和 W_0 的各层都搜索之后，再看 U_1 与 W_0 之间是否有边，如果有就添边，并且不再向 W_0 的里层继续搜索，如果没有边，则再看 U_1 与 W_1 之间是否有边. 如果有就添边，并且不再向 W_1 的里层继续搜索，依此类推，直到 U_0 各层与 W_0 各层都搜索完毕.

由于采用贪婪式搜索，可能出现不满足 1.1 节压缩图所需条件(1)的情况. 接下来，需要对压缩边进行修正. 如果存在集合 $S\subset V$，使得 $U_0\cap S\neq\varnothing, U_0\not\subset S, S\not\subset U_0$ 并且 $W_0\cap S\neq\varnothing, W_0\not\subset S, S\not\subset W_0$，则这种结构最容易产生相交的压缩边. 对这种结构，需要判断当前生成的压缩边集合 L 中是否同时出现边 $(U_0, W_0\cap S)$ 和 $(W_0, U_0\cap S)$. 如果出现，就不符合压缩图所需满足的条件(1)，需要对边进行修正. 也就是，把原来的边更新成 $(U_0\setminus S, W_0\cap S)$ 和 $(U_0\cap S, W_0)$（或者 $(U_0\cap S, W_0\setminus S)$ 和 $(U_0, W_0\cap S)$），同时把原来的压缩结点 U_0 和 W_0 分解成 $U_0\setminus S, U_0\cap S, W_0$（或者 $W_0\setminus S$，U_0，$W_0\cap S$）.

从压缩图生成过程可见，压缩结点的生成基于原始图的拓扑结构，保留了原始图中结点的信息；同时压缩边能够表示原始图的所有边，同样没有丢失任何信息. 因此压缩图是对原始图的无损压缩.

2　CLEAR 算法

给定压缩图 $G'=(V', E')$，CLEAR 算法无需解压缩，可以直接在压缩图上发现重叠社区. 算法由 3 个主要步骤构成：①选取若干压缩结点作为“种子”；② 不断优化社区适应度函数将“种子”扩展为社区；③将相似度很高的社区进行合并，得到最终的重叠社区结果. 以下，分别对这 3 个步骤加以详细描述.

2.1　“种子”选取

社会网络中的社区大多是以一个或几个点为中心的. 因此，一个最直观的想法是选取一个或几个结点作为初始社区，以其作为“种子”不断扩充.

首先，涉及的是选取谁作为“种子”的问题. “派系”是一个完全子图. 如 1.2 节所述，“派系”结构代表定义最严格的一种社区结构. “极大派系”是指不被网络中任何其他派系所包含的派系，与压缩图中的“派系”结构对应. 因此，CLEAR 算法选取“极大派系”作为“种子”，大大节省了寻找“种子”的时间.

另外，选取的“种子”大小也是一个需要考虑的问题. 也就是说，“极大派系”中包含的结点个数，也应满足一定要求. 设“种子”的大小至少为 k，则 k 既不能太大，也不能太小. 如果 k 太大，有些中小规模的社区结构不能被发现，因为这些社区结构中并不包含大小为 k 的种子，这种情况被称为“漏报”；反之，如果 k 太小，则有些种子所在的网络区域即使根本不存在社区结构，算法却仍然会对其进行扩展，这种情况称为“误报”. 通过实验发现，参数 k 选择默认值为 4 时比较适宜.

2.2　“种子”扩展

“种子”扩展涉及一个非常重要的概念，即社区适应度函数.

定义 4　给定函数 $F: S\to R$，函数 F 将原始图 G 中的某个子图 S 映射到实数域 R 上的某个值. 该值反映了 S 成为社区的程度. 值越大，说明 S 越符合社区的结构. 称函数 F 为社区适应度函数.

社区适应度函数对扩展“种子”成为社区非常重要. 记 G 的某子图 S 中所有结点构成的集合为 C. 如果集合 C 中去掉任何一个元素都不会使 S 的社区适应度函数 F 变大，且将 S 的任一邻居结点添到集合 C 中也不会使 S 的社区适应度函数 F 变大，则 S 是一个社区. 具体来说，本文采用 Lancichinetti 等[10] 提出的社区适应度函数，如公式(3)所示.

$$F=\frac{k_{\text{in}}^{s}}{(k_{\text{in}}^{s}+k_{\text{out}}^{s})^{\alpha}}. \tag{3}$$

其中：k_{in}^{s} 是指 S 中所有结点在 S 内的度数之和，其值等于所有落在 S 内的边数的两倍；k_{out}^{s} 是指 S 中所有结点在 S 外的度数之和，其值等于只有一个端点落在 S 内的边数之和. 参数 α 是正实数，用来控制社区规模.

以图 1 为例，“种子”扩展的具体步骤可描述为：①对 S 的每个邻居结点 v（图 1 中的阴影结点），计算 v 对 S 的贡献值，也就是，加入 v 后，S 的社区适应度改变量；②选择对 S 贡献最大的结点 $v_{\max}$；③ 如果 $v_{\max}$ 对 S 的贡献值为正，将其加入

S 并返回①,否则停止扩展并返回 S.

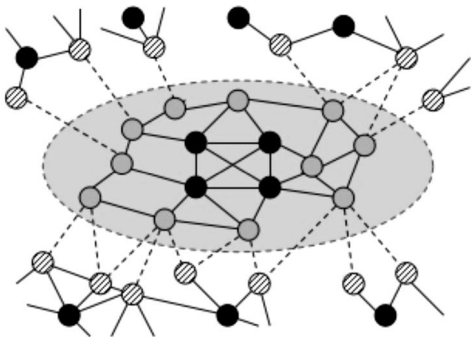

图 1 种子扩展的过程

Fig. 1 The process of seeds expanding

扩展过程类似雨滴落水,是一种从中心向周围慢慢扩散的动态场景. 对每一个“种子”都用这种方式进行扩展,当两个种子相互外扩至重叠时,就找到了两个社区的重叠结点. 扩展中,社区适应度函数和种子的选取不是固定的,可以根据情况采用不同的社区适应度函数以及不同的种子选取方法.

2.3 社区合并

在将“种子”扩展为社区的过程中,可能会出现社区间重叠度过高,甚至相互包含的情况. 考虑这两种情况,可以将重叠度过高或者完全重合的社区进行合并. 一种最简单的重叠度度量方法是计算两个社区重叠结点的个数与两个社区总结点个数的比值,如公式(4)所示:

$$\delta(S,S') = \frac{|S \cap S'|}{|S \cup S'|}. \tag{4}$$

其中,S 和 S'分别代表两个社区.

进一步,为了避免两个社区中结点个数的悬殊差异对重叠度的影响,公式(4)可以调整为

$$\delta(S,S') = \frac{|S \cap S'|}{\mathrm{Min}(|S|,|S'|)}. \tag{5}$$

其中,$\mathrm{Min}(|S|,|S'|)$代表社区 S 和 S'中结点个数最少的社区. 这样,当重叠度为 1 时,就对应社区间相互包含的情况. 因此,式(5)所示的重叠度计算公式可以更好地刻画社区之间的重叠程度.

合并过程中,通过定义阈值 e,当社区 S 和 S'的重叠度 $\delta(S,S) \geqslant e$ 时,将社区 S 和 S'进行合并. 合并所有重叠度大于阈值 e 的社区后,即可得到网络中的最终社区结果. 其中,被多个社区包含的结点为重叠结点,包含重叠结点的社区即为重叠社区.

3 实验及结果分析

为验证 CLEAR 算法的有效性和效率,对 CLEAR 算法进行分析,算法由 C++编写,所有实验均在惠普主频 2.33 GHz,4 GB 内存的 PC 上运行完成,操作系统为 Windows 7. 对比算法为三个经典的社区发现算法 LFM, LinkSCAN 和 CPM.

表 1 真实数据集

Table 1 The real datasets

真实数据集	#nodes	#links	< C >
Facebook	4 039	88 234	0.605 5
Amazon	334 863	925 872	0.396 7

表 1 总结了本文所用的两个真实数据集的信息. 其中,#nodes 代表数据中的结点数量,#links 代表数据中的边数量, < C > 是平均集聚系数. 这两个真实数据集均可从文献[11]下载.

本节基于的指标是重叠社区模块度 M^{ov}[12]. 它是在不知道真实社区结构的前提下,衡量社区划分效果的一种评价标准. M^{ov} 取值在[-1,1]之间. 值越大,社区划分效果越好.

在“Facebook”和“Amazon”这两个真实数据集上的 M^{ov} 值比较结果如图 2 所示. 可以看出,通过 CLEAR 算法得到的社区结构的模块度最高,也就是说,对于不同规模的真实网络 CLEAR 算法都有效. 由于 CPM 算法对于社区结构的定义过于严格,导致算法准确度不是很高. 而LinkSCAN

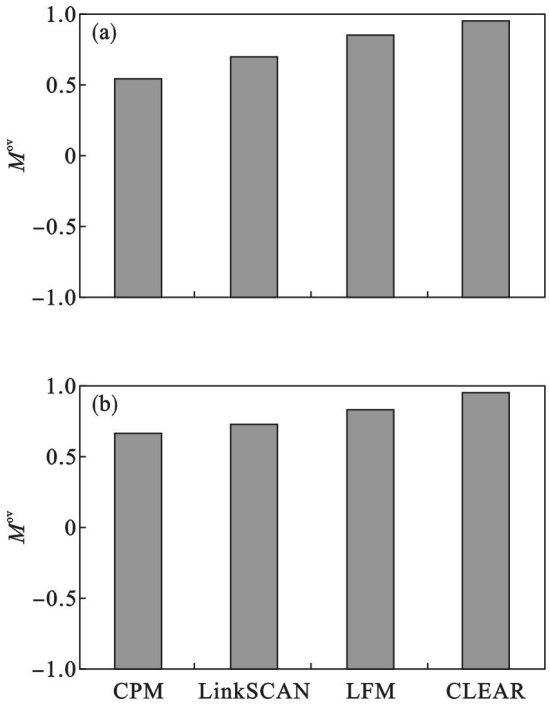

图 2 有效性的比较

Fig. 2 The comparison of effectiveness

(a)—Facebook; (b)—Amazon.

算法容易产生过多没有必要的小规模社区，使得重叠社区数量过多.

4　结　　语

重叠社区发现对研究真实网络的拓扑结构具有重要指导意义，是社区发现中一个新兴的热点. 大多数现有的重叠社区发现算法计算效率偏低，只能作用于规模较小的网络. 针对这种情况，本文提出了一种基于图压缩的重叠社区发现方法 CLEAR，能在保证准确度的情况下，大大提高单机上处理的网络规模，使有效解决大规模网络上的重叠社区发现问题成为可能. 实验对算法性能进行了验证. 结果表明，CLEAR 算法的性能优于另 3 个作为对比的经典社区发现算法.

参考文献：

[1] Aggarwal C C. Social network data analytics [M]. Berlin: Springer-Verlag, 2011: 1 – 30.

[2] Gergely P, Imre D, Illes F, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. *Nature*, 2005, 435: 814 – 818.

[3] Lim S, Ryu S, Kwon S, et al. LinkSCAN: overlapping community detection using the link-space transformation [C]// IEEE 30th International Conference on Data Engineering. Chicago, 2014: 292 – 303.

[4] Cui W, Xiao Y, Wang H, et al. Online search of overlapping communities [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. New York, 2013: 277 – 288.

[5] 陈东明，徐晓伟. 一种基于广度优先搜索的社区发现方法 [J]. 东北大学学报：自然科学版，2010，31(3)：346 – 349.
(Chen Dong-ming, Xu Xiao-wei. A community discovery method based on breadth-first search [J]. *Journal of Northeastern University: Natural Science*, 2010, 31(3): 346 – 349.)

[6] Fortunato S. Community detection in graphs [J]. *Physics Reports*, 2010, 486(3/4/5): 75 – 174.

[7] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks [J]. *Physics A: Statistical Mechanics and its Application*, 2009, 388(8): 1706 – 1712.

[8] Lee C, Reid F, McDaid A, et al. Detecting highly overlapping community structure by greedy clique expansion [C]// The 4th SNA – KDD Workshop on Social Network Mining and Analysis. Washington D C, 2010: 1 – 10.

[9] Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities [J]. *Physical Review E*, 2009, 80(1): 100 – 105.

[10] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure of complex networks [J]. *New Journal of Physics*, 2009, 11(3): 15 – 33.

[11] Leskover J. Stanford large network dataset collection [EB/OL]. (2014 – 06 – 01) [2015 – 08 – 14]. http://snap.stanford.edu/data.

[12] Lázár A, Ábel D, Vicsek T. Modularity measures of networks with overlapping communities [J]. *Europhysics Letters*, 2010, 90(1): 18 – 31.

(上接第 1538 页)

[5] Ding H, Lv J J. Comparison study of two commonly used methods for envelope fitting of empirical mode decomposition [C]//International Congress on Image and Signal Processing. Piscataway: IEEE, 2012: 1875 – 1878.

[6] Chu P C, Fan C W, Huang N. Derivative-optimized empirical mode decomposition for the Hilbert-Huang transform [J]. *Journal of Computational and Applied Mathematics*, 2014, 259: 57 – 64.

[7] Chen Q H, Huang N, Riemenschneider S, et al. A B-spline approach for empirical mode decompositions [J]. *Advances in Computational Mathematics*, 2006, 24(1/2/3/4): 173 – 175.

[8] Qin S R, Zhong Y M. A new envelope algorithm of Hilbert-Huang transform [J]. *Mechanical Systems and Signal Processing*, 2006, 20(8): 1941 – 1952.

[9] Xu Z G, Huang B X, Li K W. An alternative envelope approach for empirical mode decomposition [J]. *Digital Signal Processing*, 2010, 20(1): 77 – 84.

[10] 朱伟芳，赵鹤鸣，陈小平. 一种最小长度约束的 EMD 包络拟合方法 [J]. 电子学报，2012，40(9)：1909 – 1912.
(Zhu Wei-fang, Zhao He-ming, Chen Xiao-ping. A least length constrained envelope approach for EMD [J]. *Acta Electronica Sinica*, 2012, 40(9): 1909 – 1912.)

[11] Higham D J. Monotonic piecewise cubic interpolation, with applications to ODE plotting [J]. *Journal of Computational and Applied Mathematics*, 1992, 39(3): 287 – 294.