

关系数据库中社区发现方法研究

张 璋,张 然,朱东生

(长沙理工大学 计算机与通信工程学院,湖南 长沙 410073)

摘 要:文中在研究了现有社区发现算法的基础上,提出了一种简单的加权网络中社区发现方法。文中基于社区结构最为普遍的性质,受社会网络中真实社区结构和并行计算的任务划分规则的启发,提出了基于核心边的加权网络中社区发现方法。该方法首先依据网络中边的权值寻找核心边;然后依据相似性度量,发现网络中的一个初始社区;最后通过隶属度度量,将发现的初始社区逐步扩展成网络中的社区结构。该方法在进行社区结构发现的过程中,仅仅依赖节点所处位置的局部信息,可以在对网络进行广度优先遍历的过程中完成社区发现工作。因此该方法具有较低的计算复杂度,可以适用于大规模网络中的社区发现。通过有效性实验和效率实验,表明该方法可以有效发现大规模网络中的社区结构。

关键词:社区发现;加权网络;分布式数据库

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2014)08-0108-04

doi:10.3969/j.issn.1673-629X.2014.08.025

Research on Community Discovery Methods in Relational Database

ZHANG Zhang,ZHANG Ran,ZHU Dong-sheng

(College of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410073, China)

Abstract: A simple weighted community discovery algorithm is proposed based on study of existing community discovery methods. Based on the most commonly nature of community structures, inspired by task division rules of the real community structure and parallel computing in social networks, propose the weighted network community discovery methods based on core edge. In this method, based on the weight of edges in the network, look for the core edge; then in accordance with similarity measure, find an initial community in the network; finally through membership metrics, you will find the initial community gradually expands into community structures in a network. This method during the discovery process of community structures, relies solely on the local node location information, the network can be a breadth-first traversal of the community discovery process to complete the work. Therefore, the method has a low computational complexity, which can be applied in a large-scale network community discovery. Through experimental test on effectiveness and efficiency, the method can effectively detect large-scale network of community structures.

Key words: community discovery; weighted networks; distributed database

0 引言

随着移动通信网络^[1]的普及,社交网站的流行,从现实中获取的数据越来越庞大,数据之间的关系越来越复杂。使用传统的分析方法对这些数据进行分析处理,难以满足人们的需求。随着社区发现研究的兴起,为人们研究这些数据提供了有效的方法。随着分布式数据库的普及,数据库可以提供的功能越来越强大,计算能力也不断提升。

从现实世界中抽象得到的复杂网络,其规模都是极其庞大的,描述这些真实系统的网络,一般都是含有

一定的权值的,这些权值可能是整数,也可能是小数,甚至是负数。如何在这些加权的网络中找到其中隐含的社区依然是社区发现课题所面临的挑战。

1 相关研究

1.1 基于相似度的方法

社区可以看作是那些具有相似对象的集合^[2]。基于这种观点,很多研究人员从相似度的角度对社区发现方法进行研究,根据节点间的相似度或者相异度,不论两个节点是否有边直接相连,它们都会被划分到最

收稿日期:2013-10-12

修回日期:2014-01-25

网络出版时间:2014-05-21

基金项目:国家“863”高技术发展计划项目(2013AA01A212)

作者简介:张 璋(1987-),男,硕士研究生,研究方向为关系数据库。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140524.2150.045.html>

相似的集群中^[3]。

如果网络 G 中的节点可以映射到一个 n 维的欧式空间中,那么每一个节点都可以用它在欧式空间中的位置即坐标来表示,通过节点所在位置的坐标,可以计算任意两个节点之间的距离,也就可以知道两个节点的相异度。例如节点 u 的坐标为 (v_1, v_2, \dots, v_n) , 节点 v 的坐标为 (w_1, w_2, \dots, w_n) , 通过计算它们之间的 m 范数,得出它们的相异度^[4]。

1.2 基于模块的方法

网络中发现了社区结构之后^[5],评价所发现的社区结构的质量,最简单的方法是把发现的社区结构和网络中实际的社区结构进行对比,越符合实际网络中的社区结构,那么算法发现的社区结构越理想。但是对于大规模的网络来说,里面的社区结构是未知的,包括社区结构的个数是未知的,各个社区结构的规模也是未知的。为了评价算法所发现的社区结构的好坏,人们提出了模块度的概念^[6]。

自从提出了模块度的概念之后,把优化模块度作为进行社区发现的方法^[7]也成为了社区发现方法中的一类方法。这类方法把模块度达到最大取值作为目标函数^[8],使用组合优化的方法进行不断的尝试,最终使得模块度达到最大值或者局部的最大值。

基于网络划分的方法,研究人员根据各自不同的需要设计使用了各式各样的模块度计算方法^[9],但是应用最为广泛的还是由 Newman 和 Girvan 所提出的 NG 模块度。但是 NG 模块度有一个局限性,当网络中社区规模差距较大时,模块度最大时得到的社区并不一定是最好的。

1.3 加权网络中的社区发现

在社区结构研究的早期,社区发现^[10]是作为网络划分的一种进行研究的,因此将一些网络划分引入到了社区发现的过程中,例如 K-L 算法、谱分析方法等等。把网络划分的方法应用到社区发现中有着许多的局限性:

(1) 网络划分^[11]中,要求知道把网络划分成几个部分,而在进行社区发现的过程中,网络中所包含的社区结构的个数是未知的,也就不可能指定社区的个数。

(2) 网络划分中,要求划分之后各个区域的规模是均衡的,然而实际的网络中,社区的规模是不同的,有的社区规模很大,有的社区规模却很小。

(3) 网络划分中,要求划分后割边最小,也就是各个区域之间的联系最小,但是并不考虑各个区域内部连边的情况。

因此虽然把网络划分引入到了社区发现中,但是并不适合实际应用。现在的社区发现算法也都脱离了网络划分方法的基本方式,形成了专门针对复杂网络

中社区发现的一个分支^[12]。

2 加权网络中社区发现方法

2.1 问题引入

对有向网络,加权网络^[13]的研究目前并不多,比较经典的处理加权网络的方法是由 Newman 在 2004 年提出的。他把网络中边的权值看作是节点之间的重边,通过不断移除边介数最大的边,来得到网络中的社区结构^[14]。这种方法只能处理那些具有正整数权值的网络,如果网络中的权值是小数则无法进行处理。

对加权网络的另外一种处理方式是简单地忽略网络中边的权值,即把加权网络当作普通的无权网络进行处理。这种处理方式会丢失原网络中的许多重要信息,例如在社交网络中,边的权值可能代表了两个对象的熟悉程度;在并行计算中,边的权值可能代表了两个任务需要进行通信的数据量等等。如果简单地忽略这些权值则不能够体现这些重要的性质。

文中受到社交网络中对象间的互动关系以及并行计算中任务划分规则的启发,根据社区结构最基本的性质,即社区内部连接紧密而社区间连接稀疏,提出了基于核心边的社区发现方法。该方法首先找到网络中的核心边,然后通过相似度得到一个初步的社区结构,之后使用加权隶属度对社区的邻居节点进行处理,最终得到网络中的社区结构。文中以分布式数据库为基础,充分利用分布式数据库的高性能,实现庞大复杂网络中的社区发现。

2.2 相关定义

在社会网络中边的权值可以代表两个对象之间的交互频数,交互越频繁的两个对象属于同一个社区的可能性越大;在并行计算任务划分的网络中,边的权值可以代表两个子任务之间的通信量,通信量大的子任务应该被划分到同一个计算节点中,此外还有很多类似的现象。从这些现象中可以发现具有较大权值的边在社区结构的发现中可能有着重要的作用。为此提出了核心边的概念:

定义 1: 核心边加权网络 $G=(V, E, W)$ 中,称具有最大权值的边为网络 G 中的核心边,简称核心边。

以社会网络为例,如果两个对象,他们大部分的社交精力都用在了彼此的交互,或者他们和共同的其他对象进行交互,那么他们属于同一个社区的可能性也会很大。基于这种事实,提出了一种加权网络中简单的相似度计算方法,为了和无权网络中的相似度进行区分,文中称其为加权相似度。

定义 2: 密集社区在加权网络 $G=(V, E, W)$ 中,核心边 (u, v) 所对应节点 u 和节点 v 以及它们的邻居节点中,加权相似度不低于设定阈值的节点所形成的节

点集称为核心边所在密集社区,简称密集社区。

定义 3:社区的邻居在社区 C 外部的节点中和社区内部节点有边直接相连的节点,称为社区的邻居节点,记为 $\sigma(C)$ 。

2.3 算法简介

基于上述的分析,本节提出了基于核心边的加权网络中社区发现方法(Core Edge in Community Detec-

tion,CECD)。该方法的主要思想如图 1 所示,(a)显示的是一个未进行处理的加权网络;(b)显示的是发现网络中的核心边;(c)显示的是通过核心边而得到的一个密集社区;(d)显示的是最终得到加权网络中的社区结构。对该网络进行社区发现后得到了两个社区,它们之间连边的权值为 1。

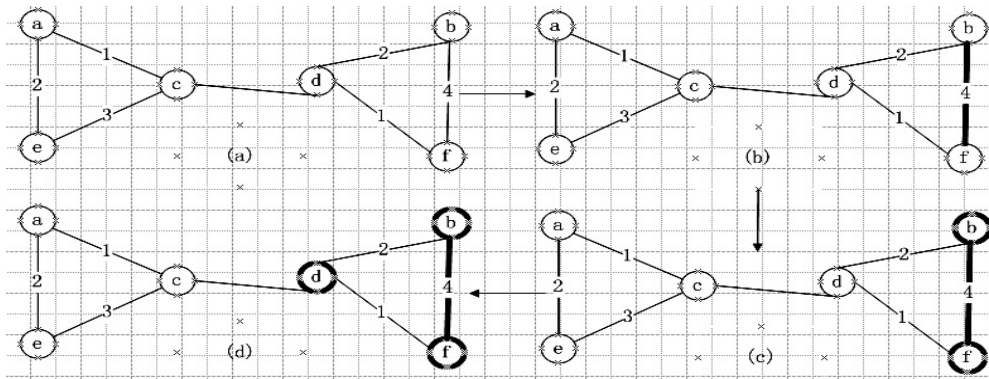


图 1 算法主要流程

算法中首先是找到网络中的核心边,然后根据核心边所对应的两个节点找到一个包含核心边的密集社区,最后通过计算隶属度将密集社区逐步地扩展为一个社区结构。

算法:

输入:加权网络 G , 加权相似度阈值, 加权隶属度阈值;

输出:加权网络 G 中的社区结构。

```

1: for  $e$  in  $V$  loop
2:    $qv.push(e)$ ;
3: end for;
4: while not  $qv.empty()$  loop
5:   find  $\max(w(u, v))$ ;
6:    $stack[u].push(v)$ ;
7:    $qv.pop(u)$ ;
8:    $qu.pop(v)$ ;
9:   for  $e$  in  $\sigma(v, w)$  loop
10:    if  $S(e, u)$  or  $S(e, v)$  then
11:       $stack[u].push(e)$ ;
12:       $qv.pop(e)$ ;
13:    end if;
14:   end loop;
15:    $flag = true$ 
16:   while  $flag$  loop
17:      $flag = false$ ;
18:     for  $e$  in  $\sigma(bd[v])$  loop
19:       if  $B(stack[u], e)$  then
20:          $stack[u].push(e)$ ;

```

```

21:    $qv.pop(e)$ ;
22:    $flag = true$ ;
23: end if
24: end loop;
25: update( $\sigma(bd[v])$ );
26: if  $\sigma(bd[v]) = \emptyset$  then
27:    $flag = false$ 
28: end if
29: end loop;
30: end loop

```

在无权网络中,可以认为所有的边的权值都相同。因此在无权网络中,不存在核心边,但是复杂网络中,节点的度分布符合幂率分布,因此在无权网络中存在度很大的核心节点。所以 CEDC 算法可以很方便地推广到无权网络中。

3 实验评估

为了评价 CVCD 算法的有效性,首先使用人工模拟数据进行有效性检测,模拟网络中含有 24 个节点,包含了四个社区,每个社区都含有 6 个节点。具体的网络结构以及划分结果如图 2 所示。图中使用不同的颜色来表示不同的社区结构,CVCD 算法发现的社区结构和网络中真实社区结构相吻合。

图 3 显示了 CEDC 算法对经典的 Zachary 空手道俱乐部进行社区发现的结果。

从图中可以看到,Zachary 网络被划分出了两个社区结构,一个以节点 1 为核心,一个以节点 34 为核心,在这个结果中,只有节点 3 和实际的社区不相符合。

但是从图中可以看出节点3和两个社区都有4条边相连,因此把节点3划分到节点34所在的社区也是合理的。这里会把节点3划分到节点34所在的社区是因为节点34比节点1具有更高的度数,CECD算法在进行社区发现时,优先发现节点34所在的社区。

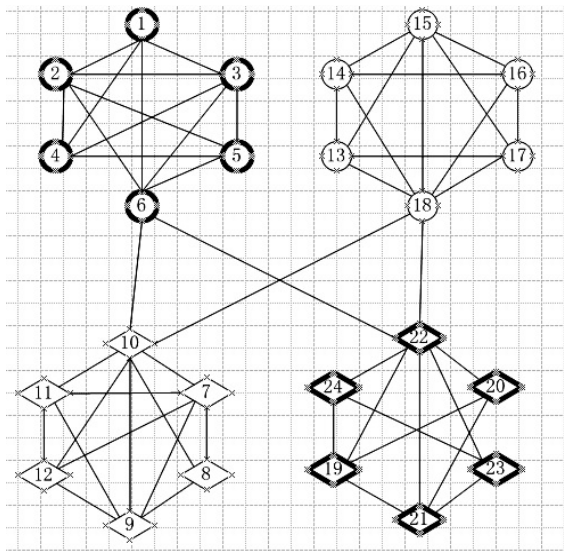


图2 人工网络中社区发现的结果

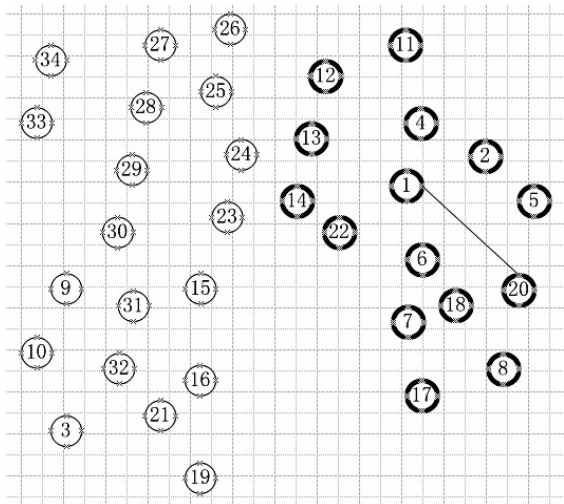


图3 Zachary 空手道俱乐部网络社区发现结果

在进行效率评估的实验中,使用的是某通信公司四月份的短信通信记录。该数据所抽象的网络总共含有386 430个节点,1 229 010条边。为了显示出算法的运行效率和网络规模的关系,实验中最初从所给的记录里抽取了100个节点,然后每次实验后再重新多抽取300个节点进行实验,最后一次实验中节点总数达到了2 700个,如图4所示。

4 结束语

通过对社交网络中真实社区的研究,以及对并行任务执行过程中任务分配原则的研究,文中提出了通过核心边,以相似度和隶属度为依据,一种简单的加权

网络中的社区发现方法。该方法只依赖于节点所在位置的局部信息,可以在对网络进行广度优先遍历的过程中完成社区的发现工作。文中对该方法进行了详细的介绍,并通过实验验证了该方法的有效性,以及可以应用到大规模无标度网络的适用性。

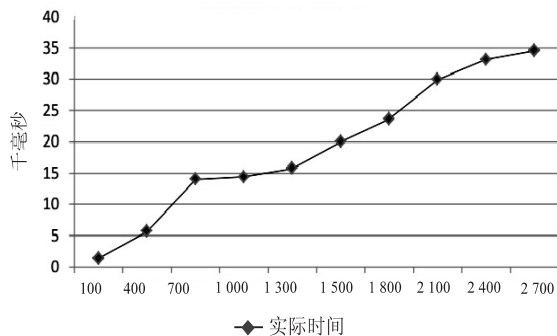


图4 CECD算法运行时间与节点规模关系图

在社区结构的发现中,文中对加权网络进行了深入的研究,但是没有涉及到负数权值网络,如何在负数权值的网络中发现其中隐含的社区结构依然是社区结构发现这一课题所面临的严峻的挑战。

参考文献:

- [1] Shekhar S, Lu Chang-Tien, Zhang Pusheng. A unified approach to detecting spatial outliers[J]. GeoInformation, 2003, 7(2): 139-166.
- [2] Liu R Y, Parelius J M, Singh K. Multivariate analysis by data depth: descriptive statistics, graphics and inference[J]. Annals of Statistics, 1999, 27(3): 783-858.
- [3] Giannella C, Han Jiawei, Pei Jian, et al. Mining frequent patterns in data streams at multiple time granularities[M]//Next generation data mining. [s. l.]: [s. n.], 2003.
- [4] Han Jiawei, Kamber M. Data mining: concepts and techniques [M]. Beijing: Higher Education Press, 2001.
- [5] Hand D, Mannila H, Smyth P. Principles of data mining[M]. Beijing: China Machine Press, 2003.
- [6] Dunkel B, Soparkar N. Data organization and access for efficient data mining[C]//Proc of the 15th international conference on data engineering. Sydney: IEEE, 1999: 522-529.
- [7] Han Jiawei, Pei Jian, Yin Yiwen. Mining frequent patterns without candidate generation[C]//Proc of ACM-SIGMOD international conference on management of data. Dallas, TX: [s. n.], 2000: 1-12.
- [8] David C, Han Jiawei, Vincent T N, et al. Maintenance of discovered association rules in large databases: an incremental updating technique[C]//Proc of the 12th international conference on data engineering. New Orleans, Louisiana, USA: [s. n.], 1996.
- [9] Tan Pangning, Steinbach M, Kumar V. Introduction to data mining[M]. Beijing: Post & Telecom Press, 2006.

务资源预留,可以在 Web 服务失效时直接进行服务替换,有效降低了 Web 服务调整的时间。存在的问题是预留服务资源会造成一定的资源浪费和资源维护费用。但通过服务预留建立服务备份集仍然是当前 Web 服务调整的主要方法。

基于合同网的 Web 服务调整和基于分层结构的 Web 服务调整都是通过建立服务备份集的方式实现快速服务故障恢复的。与基于服务预留的 Web 服务调整所不同的是,这两种方法都建立了组合服务完整的备份集用于服务故障恢复。同时,基于合同网的 Web 服务调整还将合同网的熟人库、可信度等概念引入到 Web 服务调整中来,用于缩小服务搜索空间,减少服务选择时间。基于分层结构的 Web 服务调整则是通过对 Web 服务选择过程进行阶段化、层次化分解,建立不同阶段的模型和递进式的结构,使服务失效时的 Web 服务二次组合变得更加容易、快捷。

综上所述,Web 服务调整方法的综合对比分析结果如表 1 所示。

表 1 五种 Web 服务调整方法对比

	搜索空间	空间耗费	避免“二次失效”	故障恢复时间
影响区域	大	小	无	较长
可信度	大	小	有	中等
服务预留	大	较大	无	短
合同网	小	大	有	最短
分层结构	小	大	无	短

从综合分析来看,基于合同网的 Web 服务调整策略在有一定的空间耗费(服务预留)的情况下,综合性能相对较优。

4 结束语

文中针对 Web 服务调整这一热点问题,简要回顾了国内外的研究现状,并且对比分析了这些方法的优缺点。对比发现,建立服务备份集是一种应用多、效果好的方法,同时,通过引入阶段化分析方法、可信度评估方法等都能够很好地改善组合服务的故障恢复性能,这也为后续进行 Web 服务调整策略的改进提供了基础。

参考文献:

[1] Hirschfeld R,Kawamura K. Dynamic service adaptation[C]//Proceedings of 24th international conference on distributed computing systems workshops. Tokyo:IEEE,2004:290-297.

[2] Moser O,Rosenberg F,Dustdar S. Non-intrusive monitoring and service adaptation for WS-BPEL[C]//Proceedings of the 17th international conference on World Wide Web. Beijing: IW3C2,2008:815-824.

[3] 刘伟,鱼滨. 基于 QoS 的动态服务组合研究[J]. 计算机技术与发展,2007,17(5):140-143.

[4] 罗森,许斌,孙科武. 一种基于服务的自适应网络应用框架[J]. 小型微型计算机系统,2013,34(1):16-22.

[5] 魏乐,赵秋云,舒红平. 云制造环境下基于 QoS 的组合云服务自适应调整[J]. 兰州大学学报(自然科学版),2012,48(4):98-104.

[6] Fang Kun,Li Jianxin,Sun Hailong,et al. Strategy-based two-level fault handling mechanism for composite service[C]//Proc of IEEE 2nd international conference on software engineering and service science. Beijing:IEEE,2011:494-499.

[7] 宋巍,唐金辉,张功萱,等. WS-BPEL 服务可替换性分析[J]. 中国科学:信息科学,2012,42(3):264-279.

[8] Hass H,Bmwn A. Web services glossary[EB/OL]. 2010. <http://www.w3.org/TR/ws-gloss/>.

[9] Papazoglou P M. Web 服务原理和技术[M]. 龚玲,张云涛,译. 北京:机械工业出版社,2010.

[10] 翟岩龙. 开放网络环境中动态自适应服务组合技术研究[D]. 北京:北京理工大学,2009.

[11] Yu Tao,Lin Kwei-Jay. Service selection algorithms for Web services with end-to-end QoS constraints[C]//Proc of 2004 IEEE international conference on e-commerce technology. [s. l.]:IEEE,2004:129-136.

[12] 张跃. 基于可信度的组合服务自适应维护方法研究[D]. 沈阳:辽宁大学,2011.

[13] 伍章俊. 云 workflow 服务组合与活动调度策略研究[D]. 合肥:合肥工业大学,2011.

[14] 秦胜君. 复杂适应信息系统体系结构的研究与应用[D]. 大连:大连海事大学,2011.

[15] Chafle G, Dasqputa K, Kumar A,et al. Adaptation in Web Service composition and execution[C]//Proc of IEEE international conference on Web Service. Chicago:IEEE,2006:549-557.

(上接第 111 页)

[10] 黄德才,张良燕,龚卫华,等. 一种改进的关联规则增量式更新算法[J]. 计算机工程,2008,34(10):38-39.

[11] 商志会,陶树平. 一种高效的关联规则增量更新算法[J]. 计算机应用,2005,25(4):830-832.

[12] Zhu Honglei,Li Ming. An incremental updating algorithm for maintaining discovered association rules[J]. Application Re-

search of Computer,2004(9):107-109.

[13] Peng D,Dabek F. Large-scale incremental processing using distributed transactions and notifications[C]//Proc of OSDI. [s. l.]:[s. n.],2010.

[14] Cand O. Nova;continuous pig/hadoop workflows[C]//Proc of SIGMOD. [s. l.]:[s. n.],2011.

关系数据库中社区发现方法研究

作者: [张璋](#), [张然](#), [朱东生](#), [ZHANG Zhang](#), [ZHANG Ran](#), [ZHU Dong-sheng](#)
作者单位: [长沙理工大学 计算机与通信工程学院, 湖南 长沙, 410073](#)
刊名: [计算机技术与发展](#) 
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2014(8)

本文链接: http://d.wanfangdata.com.cn/Periodical_wjfz201408025.aspx