并行社区发现算法的可扩展性研究

刘强[1]，贾焰[1]，方滨兴[1,2]，周斌[1]，胡玥[1]，黄九鸣[1]
（1. 国防科技大学计算机学院，湖南 长沙 410073；2. 北京邮电大学计算机学院，北京 100876）

摘 要：社交网络中往往蕴含着大量用户和群体信息，如话题演化模式、群体聚集效应以及信息传播规律等，对这些信息的挖掘成为社交网络分析的重要任务。社交网络的群体聚集效应作为社交网络的一种特征模式，表现为社交网络的社区结构特性。社区结构的发现已成为其他社交网络分析任务的基础和关键。随着在线社交网络用户数量的急剧增长，传统的社区发现手段已经难以适应，从而催生了并行社区发现技术的发展。对当前主流并行社区发现方法 Louvain 算法和标签传播算法在超大规模数据集上的可扩展性进行了研究，指出了各自的优缺点，为后续应用提供参考。

关键词：社区发现；并行算法；可扩展性
中图分类号：TP311
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018052

Research on the scalability of parallel community detection algorithms

LIU Qiang[1], JIA Yan[1], FANG Binxing[1,2], ZHOU Bin[1], HU Yue[1], HUANG Jiuming[1]
1. College of Computer, National University of Defense Technology, Changsha 410073, China
2. College of Computer, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: The social network often contains a large amount of information about users and groups, such as topic evolution mode, group aggregation effect, the law of information dissemination and so on. The mining of these information has become an important task for social network analysis. As one characteristic of the social network, the group aggregation effect is characterized by the community structure of the social network. The discovery of community structure has become the basis and key point of other social network analysis tasks. With the rapid growth of the number of online social network users, the traditional community detection methods have been difficult to be used, which contributes to the development of parallel community detection technology. The current mainstream parallel community detection methods, including Louvain algorithm and label propagation algorithm, were tested in the large-scale data sets, and corresponding advantages and disadvantages were pointed out so as to provide useful information for later applications.

Key words: community detection, parallel algorithm, scalability

1 引言

随着 Web 2.0 技术的迅速发展，人类进入了在线社交网络时代。目前，国内社交媒体如新浪微博、腾讯微博用户数已经超过 10 亿，国外社交媒体如 Twitter、Facebook 等每月活跃用户数达到 13 亿。社交网络用户数量的急剧增长使当前的社交网络呈现出节点的大规模性、关系结构的高度复杂性以及网络的多维演化特性，这些特性对社交网络分析技术提出了新的挑战。

收稿日期：2017-01-08；**修回日期**：2017-12-02
通信作者：刘强，liuqiang1981@nudt.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2017YFB0803303）；国家自然科学基金资助项目（No.61502517, No.61472438）
Foundation Items: The National Key Research and Development Program of China (No. 2017YFB0803303), The National Natural Science Foundation of China (No.61502517, No.61472438)

社区结构发现技术作为社交网络分析的一项基本技术，对理解社交网络的结构特性、功能特性以及信息传播特性具有重要作用。当前，社区发现方法按照是否需要网络的全局结构信息可分为基于全局的社区发现方法和基于局部的社区发现方法[1]。基于全局的社区发现方法主要有基于图划分[2]、模块度优化[3,4]、随机游走[5]以及谱聚类[6]等。在这些方法中，基于模块度优化的方法及其改进是目前使用较为广泛的一种全局性社区发现方法。基于局部的社区发现方法只需要网络的局部信息。目前，主要方法有基于标签传播[7~9]、基于局部扩展优化[10,11]以及派系过滤算法[12~14]等。在局部社区发现算法中，基于标签传播的社区发现算法及其改进是一种使用较为广泛的局部社区发现方法。

然而，面对日益增长的社交网络规模以及节点之间复杂的交互关系，大多数传统的社区发现方法已经不能有效胜任社交网络的这些新变化，迫切需要能够处理大规模社交网络的高效、准确的社区发现方法。当前，相关学者已经提出了基于并行的社区发现方法来解决大规模社交网络的社区发现问题。文献[15,16]在文献[17]的基础上，提出了面向大规模网络的社区发现并行社区发现算法，该类算法利用图划分技术加速了 Louvain 算法中最耗时的第一阶段的并行，并且保证最终能够得到具有较大网络模块度的社区划分结果。Staudt 等[18]提出了一种基于共享内存的集成学习策略的并行社区发现方法，通过将 Louvain 算法与标签传播算法进行集成，提高了社区划分的准确性。Lu 等[19]在对 Louvain 算法改进的基础上，提出了一种基于多线程的并行启发式社区发现方法。相对于传统的模块度优化方法，基于标签传播的社区发现方法由于其较低的时间复杂度，在并行社区发现方面也得到了一些研究者的关注。Akshay[20]和从玉相等[21]对传统标签传播算法（LPA, label propagation algorithm）进行了改进，提出了基于 Mapreduce 的并行 LPA 社区发现方法。Zhang 等[22]提出了一种基于灰色关联分析的分布式标签传播算法。

除了上述 2 种主流的并行社区发现方法之外，近年来，还有一些其他并行社区发现方法相继被提出。Bae 等[23]基于 gossip 协议，提出了一种基于 infomap 的并行社区发现方法。李春英等[24]在提出的最小最大团标签传播算法的基础上，利用分布式计算模型实现了该算法的并行版本，并在相关真实和人工合成网络上验证了算法的有效性和可扩展性。Peng 等[25]提出了一种基于随机块模型的并行社区发现方法，并通过多阶段最大似然法对随机块模型中的参数进行确定。

虽然上述一些方法在一些百万级节点和千万级节点网络数据集上进行了实验验证，证明了其方法的合理性和正确性。然而，随着网络数据规模的持续快速增长，上述方法在超大规模网络节点数据集上的可扩展性是一个值得研究的课题。可扩展性是评估并行算法优劣的一个重要度量标准。具体来说，并行算法的可扩展性是指随着处理器数目的增加适当增加问题的规模，算法的性能能否与问题规模呈线性比例增长。本文主要选取了 2 种主流社区并行算法即 Louvain 算法和标签传播算法，对其在超大规模数据集上的可扩展性进行了研究，为后续在超大规模网络上的社区发现算法设计提供参考。

2 大规模网络并行社区发现算法

2.1 算法的基本原理

1) 基于模块度优化的 Louvain 社区发现算法[17]

模块度用于度量网络中各个社区内部节点之间联系的紧密程度。如果一个网络具有较高的模块度值，则网络中各个社区内部的节点交互较为紧密，社区之间的节点交互则较为稀疏。因此，将模块度作为一个优化函数，求得其全局最优解，即可得到一个网络的最优社区划分。

Louvain 算法的具体步骤如下。

步骤 1 首先将网络中的每个节点分派到唯一的一个社区中，然后将节点按顺序在这些社区间进行移动，并计算相应模块度的变化值，哪个变化值最大就将节点移动到相应的社区中去。按照这个方法反复迭代，直到网络中任何节点的移动都不能使网络的总模块度值变化为止。

步骤 2 将第一步得到的社区视为新节点，重新构造子图，则该子图中节点之间边的权重为节点对应的原社区之间各边的权重的总和。

步骤 3 重复执行上述步骤，直到迭代多次后网络的模块度值不再变化，此时，得到网络的最终社区划分。

2) 基于标签传播的社区发现算法[7]

基于标签传播的社区发现算法的基本思想是

通过节点标签的传播，实现将具有同一标签的网络节点划分到同一社区当中。

LPA 的具体步骤如下。

步骤 1　初始化节点的社区标签，可将网络中每个节点的编号当作该节点的初始社区标签。

步骤 2　为网络中的所有节点指定一个随机的处理顺序，按照该顺序取出相应的节点并进行处理，即将该节点按其周围邻居中节点数量最多的标签对其进行更新（如果数量相同，就随机取一个）。

步骤 3　不断迭代，执行步骤 2，直到网络中所有节点的标签不再变化，达到稳定状态。

步骤 4　将具有相同标签的节点划分到同一个社区。

2.2　2 种并行社区发现算法的基本步骤

目前的并行算法框架主要有 Hadoop 和 Spark。Hadoop 是一个能够对大量数据进行分布式处理的软件框架，具有高可靠性和高扩展性等优点，其主要缺点是仅支持 map 和 reduce 这 2 种操作；map 操作后的中间结果需要存入磁盘；任务调度和启动开销大；内存的利用不够充分；不适合迭代计算等。Spark 主要具有以下优点：丰富的 API；能够充分利用内存，减少了磁盘 I/O 的操作；比较适合于利用中间结果的迭代计算。Spark 的不足在于其较大的内存消耗。

通过上面的分析可以看到，Spark 是基于内存的迭代计算，适用于需要多次操作特定数据集的应用场合，而 Louvain 算法在执行过程中需要用到各个当前步骤的社区划分情况，然后不断进行迭代操作，进行社区的扩展和合并。因此，在并行 Louvain 算法的框架选择上，宜选用 Spark 框架提升其性能。在上面的分析中也指出，Spark 框架对内存的消耗较大，对于标签传播算法而言，该算法具有较低的时间复杂度，且在处理超大规模数据时不需要对中间结果进行迭代操作，在进行算法的可扩展性研究时，宜选用 Hadoop 并行处理框架。

1) 并行 Louvain 的基本步骤

首先，对从社交网络中获取的大规模数据进行预处理[16]：利用 Spark 平台的 Graphx 将原始网络数据加载到系统中，Graphx 识别出顶点，并将顶点和边的集合载入内存；然后，根据模块度增量计算式，构造模块度增量矩阵：将初始网络的每个节点视为一个社区，计算各个社区之间的模块度增量，并构造相应的增量矩阵；其次，不断迭代计算发现新社区：找出所有具有最大模块度增量的节点对进行合并，并更新所有社区之间的模块度增量。当最大模块度增量为负时，社区识别过程结束。

2) 并行 LPA 的基本步骤[20]

通过将网络中的节点及其邻居节点定义为一种新的 NodeModel 数据模型，将 MapReduce 化的并行标签传播算法的输入数据集转化为 NodeModel 类型的数据集合。整个过程采取局部同步和整体异步相结合的方式来实现 LPA 的 MapReduce 并行化。在基于 MapReduce 的 LPA 社区发现算法中，需要对整个网络数据进行多次迭代，直到用户的标签达到稳定状态或预先设置的收敛条件。

首先，将整个网络数据的一次完整的迭代过程分解成 n 次小迭代过程，每次小迭代只对整个网络的 $\frac{1}{n}$ 部分数据进行处理。小迭代的 map 阶段根据 map()函数处理网络的 $\frac{1}{n}$ 数据后得到的结果，即节点的标签得到更新后的结果，将在 reduce 阶段反馈到节点的邻居节点列表中，即完成节点邻居列表里节点的标签信息。一次小迭代完成后的结果将作为下次小迭代的 map 阶段的输入，接下来的小迭代将处理另一个 $\frac{1}{n}$ 数据，并将处理后的结果在 reduce 阶段反馈到邻居节点列表中。不断重复这个过程，直到网络中的所有数据都经过处理。最终，该轮整个网络数据的大迭代过程完成。

3　实验分析

本文实验用到集群的最大节点数目为 64，每个节点的软件环境配置如下：操作系统为 Ubuntu 14.04，并行计算框架 Hadoop 版本号为 2.6.0，Spark 版本号为 1.6.2。其中，有一个主节点，它的配置是 8 核处理器，64 GB 内存，500 GB 硬盘，1 Mbit/s 带宽；其余的从节点的配置是 8 核处理器，32 GB 内存，100 GB 硬盘，1 Mbit/s 带宽。

本文实验用到的测试数据集如表 1 所示，所有网络为无向不加权网络。在表 1 中，Orkut 为 Google 公司推出的社会性网络服务网站上的用户交互关系所形成的一个社交网络数据集[26]，节点的规模达到了百万级。Sina 数据集 1 和 Sina 数据集 2 为新浪公司推出的微博网站上的用户之间交互所形成的

社交网络数据集，该数据集是通过湖南蚁坊软件股份有限公司的爬虫系统获取的。2016 年 11 月至 2017 年 6 月，微博注册用户之间的转发关系网络，其中，数据集 1 为从数据集 2 中抽取的子集，数据集 1 的节点规模达到了 2 000 万，数据集 2 的节点规模达到了亿级。Twitter 为社交网络 Twitter 上由用户之间关注关系所构成的数据集[27]，节点规模达到了将近 5 000 万。以上 4 个数据集的格式如表 2 所示。在表 2 中，第 1 列和第 2 列分别对应于网络中的源节点和目标节点，每一行对应源节点和目标节点之间的连边。

表 1　　测试数据集

测试数据	节点数/个	边数/条
Orkut 数据集	3 072 441	117 185 083
Sina 数据集 1	20 000 000	269 117 041
Sina 数据集 2	115 205 577	3 504 379 868
Twitter 数据集	48 012 340	1 468 365 182

表 2　　数据集格式

源节点	目标节点
1	2
2	3
2	4
3	5
⋮	⋮

为了对并行社区发现算法的可扩展性进行研究，对相关算法采用 Java 编程语言进行了实现。其中，并行 Louvain 算法根据文献[16]的相关算法步骤进行了实现，并行 LPA 根据文献[21]的相关算法步骤进行了实现。

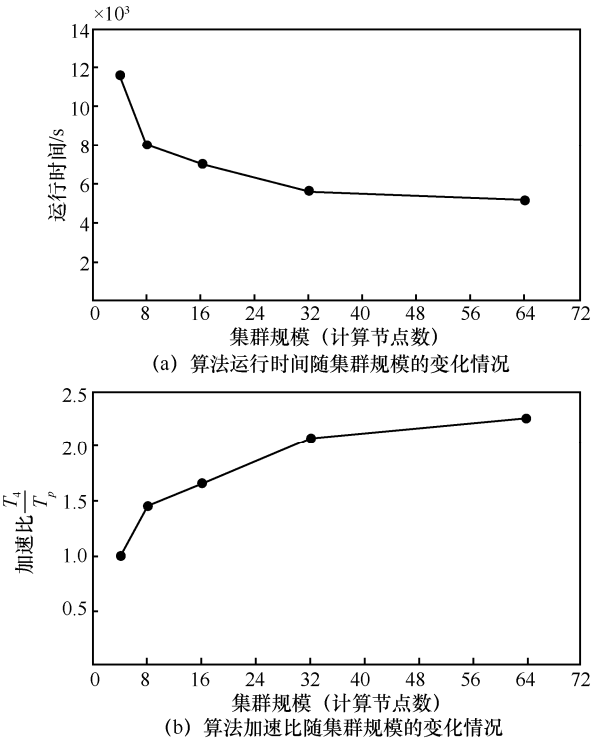
基于 Louvain 的模块度优化社区发现方法的时间复杂度为 $O(n\log n)$。其中，n 为网络中节点的个数。随着数据集规模的增加，算法时间复杂度呈 $n\log n$ 级别的增长。本文分别按照百万级节点、千万级节点和亿级节点的规模对算法的可扩展性进行了测试。

由于 Louvain 算法的时间复杂度较高，且测试数据规模较大，算法在一个计算节点上串行或在 2 个计算节点上并行的运行时间过长，于是该算法采用 4 个计算节点的 Spark 集群的运行时间作为加速比计算过程中的参考基准，故加速比定义为 $\frac{T_4}{T_p}$。

在 Louvain 并行算法执行过程中，在计算一个节点加入某个社区的模块度增量时，如果节点之间的状态信息（如社区编号）更新在同一个子任务内，那么各个子任务之间不需要产生额外的通信开销；如果节点间的状态信息更新在不同的子任务内，则节点之间需要进行通信。对于超大规模网络而言，例如，本实验中的 Sina 数据集 2 和 Twitter 数据集，其边的数目远大于节点的数目，尤其对于节点数目达到亿级别，边的规模达到几十亿级别时，每个节点都需要与相邻的节点之间进行通信，导致集群节点之间需要较大的通信开销，Louvain 算法在上述节点规模下无法在有效的计算时间内得到社区划分的最终结果，从而说明了该算法的可扩展性在较大千万级节点规模和亿级节点规模的网络数据集上存在不足，成为该社区发现算法的瓶颈。下面，对 Louvain 算法在百万级节点规模和较小千万级节点规模网络上的可扩展性进行了实验研究和分析。

图 1 和图 2 分别展示了并行 Louvain 算法在 Orkut 数据集和 Sina 数据集 1 上的可扩展性实验测试情况。由图 1(b)可知，当测试数据为百万级规模网络 Orkut 时，加速比不断增长，即使达到 32 个计算节点之后加速比仍然在增大，但是可以看到加速比增长幅度明显放缓。当测试数据为 2 000 万规模的 Sina 数据集 1 时，如图 2(b)所示，随着集群规模增大，加速比不断增加，从 32 个计算节点的集群规模开始，加速比开始出现下降趋势。另外，Louvain 算法的 Spark 并行实现版本在集群规模增加过程中，其运行时间和加速比不是一直增长的，而是会出现瓶颈，这主要是由于在大规模网络的并行算法运行过程中，计算节点间通信的时间与并行社区划分后所收益的时间相抵，甚至计算节点间通信的时间大于并行社区实现所收益的时间。

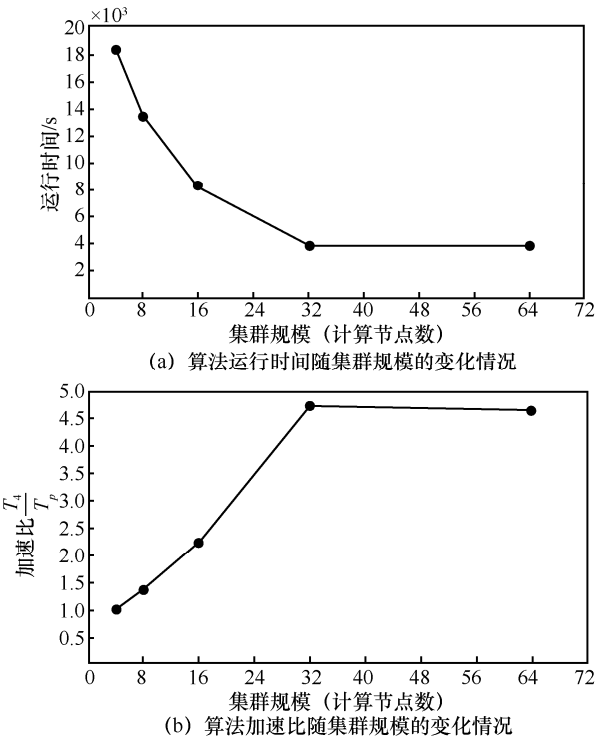
同时，研究了不同数据规模下并行 Louvain 算法运行时间及加速比的变化。图 3 展示了不同数据规模下并行 Louvain 算法运行时间分布情况。由图 3 可知，在集群规模较小时，算法运行时间在百万级节点规模以及较小千万级节点规模网络数据集上的运行时间差异显著。当集群节点为 4 和 8 时，算法运行时间的差异分别 6 637 s 和 5 407 s，集群规模从 16 开始，算法运行时间差开始放缓，甚至趋于一致。由此可见，算法在小规模集群上的运行时间对数据规模变化比较敏感。



（a）算法运行时间随集群规模的变化情况

（b）算法加速比随集群规模的变化情况

图 1　并行 Louvain 算法在 Orkut 数据集上的可扩展性测试研究



（a）算法运行时间随集群规模的变化情况

（b）算法加速比随集群规模的变化情况

图 2　并行 Louvain 算法在 Sina 数据集 1 上的可扩展性测试研究

图 4 展示了不同数据规模下并行 Louvain 算法的加速比分布情况。由图 4 可知，算法的加速比在百万级节点规模以及较小千万级节点规模网络上的加速比变化趋势基本一致，都是逐渐增长。但该算法在

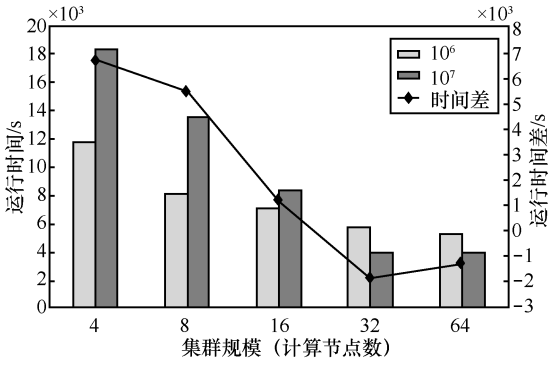


图 3　不同数据规模下并行 Louvain 算法运行时间的分布情况

较小千万级节点规模网络上的加速比相对于百万级网络来说，得到了较大的提升，这说明算法在百万级和较小规模千万级网络数据规模上的可扩展性较好。

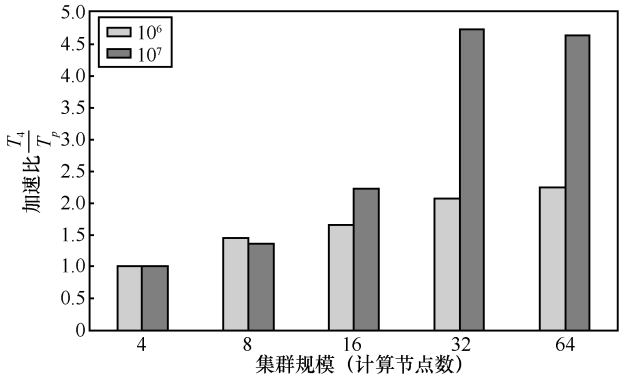
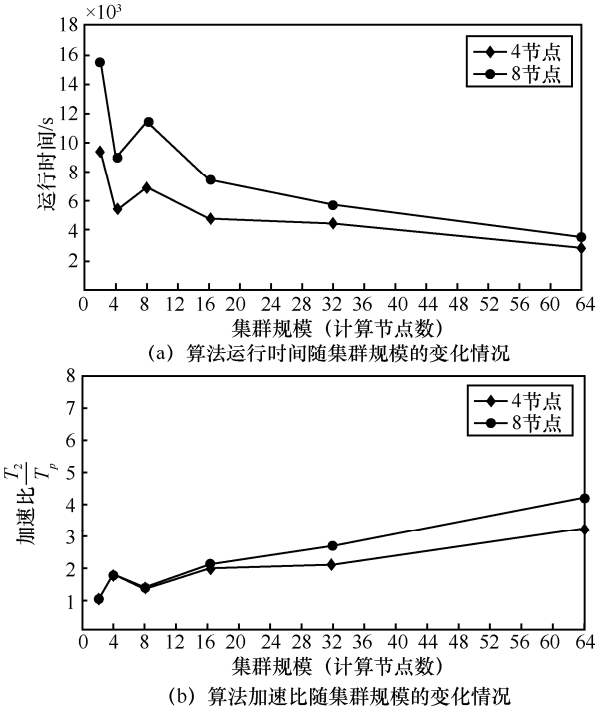


图 4　不同数据规模下并行 Louvain 算法的加速比分布情况

对于标签传播算法而言，由于其时间复杂度为 $O(n\langle k\rangle)$，其中，n 为网络中节点的个数，$\langle k\rangle$ 为网络的平均度数值。因此，标签传播算法的时间复杂度较低，尤其在稀疏网络上，可近似为 $O(n)$。本文分别在较大千万级节点规模和亿级节点规模的数据集上对并行标签传播算法的可扩展性进行了研究，并采用 2 个计算节点的 MapReduce 集群的运行时间作为加速比计算过程中的参考基准，加速比定义为 $\frac{T_2}{T_p}$。图 5 和图 6 分别展示了并行标签传播算法在 Twitter 数据集和 Sina 数据集 2 上的可扩展性测试情况。一般而言，标签传播算法的迭代次数达到一定数目后，将达到收敛，网络中节点的标签状态不再发生变化。

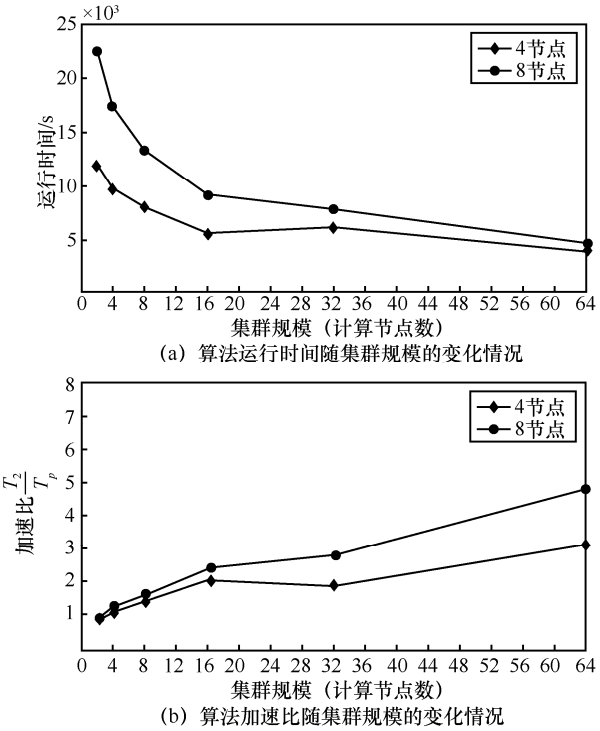
目前，相关研究表明[28]，算法收敛所需要的迭代次数独立于网络的规模，一般在迭代 4~5 次后，超过 95%的节点将归入正确的社区。本文实验分别设置算法的迭代次数为 4 次和 8 次。由图 5(a)和图 6(a)可知，



（a）算法运行时间随集群规模的变化情况

（b）算法加速比随集群规模的变化情况

图 5 并行标签传播算法在 Twitter 数据集上的可扩展性测试研究



（a）算法运行时间随集群规模的变化情况

（b）算法加速比随集群规模的变化情况

图 6 并行标签传播算法在 Sina 数据集 2 上的可扩展性测试研究

当集群规模较小时，不同迭代次数之间所花费的时间差较大，随着集群规模的增大，不同迭代次数之间所花费的时间差逐渐减小。因此，为了得到高质量的社区划分效果，可以在集群计算节点规模超过 32 时，适当地增加算法的迭代次数。由图 5(b)和图 6(b)可知，当计算节点数小于或等于 16 时，对于 Twitter 数据集和 Sina 数据集 2 而言，不同迭代次数下算法的加速比比较接近。另外，随着网络规模的增大，算法的加速比得到了一定的提升，这说明并行标签传播算法在千万级和亿级节点规模的数据集上的可扩展性较好。

另外，由图 5 可知，对于 Twitter 数据集而言，并行标签传播算法的运行时间随着集群规模的增大而减小，虽然当迭代次数为 4 和 8 时，出现小范围波动，但整体的变化趋势保持一致，特别是当计算节点数达到 64 时，运行时间基本稳定在 4 000 s 左右。对于 Sina 数据集 2，在迭代次数分别为 4 和 8 时，并行标签传播算法的运行时间随着集群规模的增加而不断减小，直到 64 个节点时趋于稳定，稳定在 5 000 s 左右。

图 7 展示了并行标签传播算法的运行时间随数据规模的分布情况。由图 7 可知，并行标签传播算法在千万级、亿级规模的网络上，随着集群规模的变化，运行时间差逐渐减小，特别地，在小规模集群上，算法的运行时间差异显著，差异最大的是在计算节点为 4 时的小集群上。由此可见，并行标签传播算法在小规模计算节点上对于数据规模较为敏感，大规模集群上对数据规模的敏感程度逐渐减小。

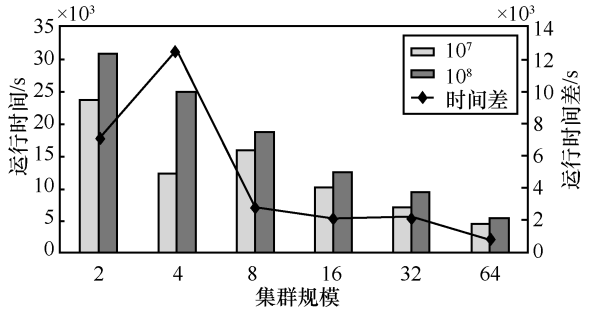


图 7 不同数据规模下并行标签传播算法的运行时间分布情况

同时，本文研究了网络的模块度随迭代次数的变化规律。在这里，以 2 个计算节点的小规模集群为例，对 Sina 数据集 2 采用并行标签传播社区发现算法进行社区划分。根据社区划分的结果以及原始网络结构，计算迭代次数分别为 4、8、12、16 时，社区划分结果的模块度值。本文以集群规模为 2 时为例，得到如图 8 所示的结果。

由图 8 可知，在迭代次数分别为 4、8、12、16 时，并行标签传播算法社区划分的模块度分别为

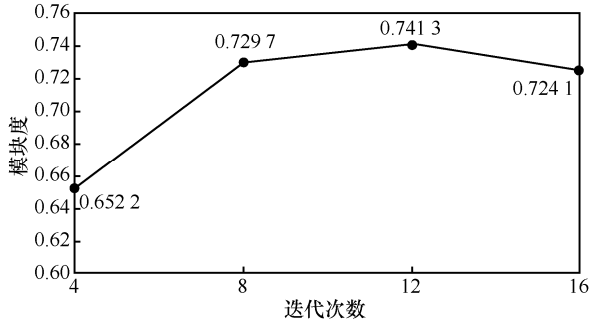


图 8　并行标签传播算法在 Sina 数据集 2 上社区划分所得的模块度随迭代次数的变化情况

0.652 2、0.729 7、0.741 3、0.724 1。模块度是用来衡量网络社区结构紧密程度的指标，可用来评估社区发现算法的优劣。对于同一个网络而言，如果某个算法在该网络上得到的社区划分结果的模块度值高于其他算法，说明该算法社区划分的效果较好。如图 8 所示，随着并行标签传播算法中迭代次数的增加，模块度并不是线性增加，而是逐渐增加，并在迭代次数大于 12 时，模块度开始减小。由此可见，并行标签传播算法可以实现快速收敛，但并不是说迭代次数越大，所划分得到的社区划分效果越好，如果迭代次数过大，可能会造成标签的误判从而降低社区划分的质量。因此，应当合理选择并行算法的迭代次数。迭代次数的增加会使算法的运行时间增长较快，从而影响算法的效率，但迭代次数过低又会降低社区划分的效果。

由上述分析可知，相对于 Louvain 算法，基于标签传播的社区发现方法具有较低的时间复杂度，能够应对超大规模网络的社区发现任务，其可扩展性在千万级和亿级节点规模的网络上也得到了较好的验证。例如，在上述实验中，对于只有 2 000 万节点的 Sina 数据集 1，在集群计算节点规模为 64 时，Louvain 算法需要 3 954 s，而对于具有 1 亿节点的 Sina 数据集 2 时，基于标签传播的社区发现方法在迭代 4 次的情况下只需要 3 815 s 即可得到社区划分结果，即使迭代 8 次，也只需要 4 600 s 左右。

对于上述 2 种主流并行社区发现算法在不同节点规模网络数据集上的实验对比分析可以发现，随着集群规模的增加，算法的运行时间和加速比的变化趋势基本一致。然而，对于并行 Louvain 算法而言，其在百万级节点规模和较小千万级节点规模网络数据集上的加速比的提升幅度存在一定的差异性，集群规模的增加在较大规模数据集上对算法的加速作用较为显著，但对于不同规模的数据集，当集群规模增大到一定数目时，算法的加速比提升缓慢，甚至会呈现一定程度的下降，这是由于较大的集群规模增加了计算节点之间的通信开销，造成了算法并行度的缩减。对于并行标签传播算法而言，算法的加速比与迭代次数有关，随着集群规模的增大，其在较大千万级节点规模和亿级节点规模网络数据集上的加速比的提升幅度较为一致。当集群规模增大到一定数目时，并行标签传播算法的加速比提升也会逐渐放缓。只需合理选择算法的迭代次数和集群规模，并行标签算法就能够完全胜任超大规模网络数据集上的社区发现任务，其可扩展性优于 Louvain 并行算法。

4　结束语

本文对当前主流并行社区发现算法的可扩展性进行了实验研究和分析，具体研究了在超大规模数据集上不同集群规模、不同数据规模变化情况下，算法的运行时间和加速比的变化情况。实验研究表明，基于模块度优化的并行 Louvain 算法在百万级节点规模和较小千万级节点规模网络上的可扩展性较好，能够较好地完成社区发现任务。当节点规模达到亿级以后，并行 Louvain 算法由于其较高的时间复杂性和集群中各节点之间较大的通信开销，不能在有效时间内得到社区划分的结果。并行标签传播算法由于其较低的时间复杂度和较低的通信开销，具有良好的可扩展性，能够较好地胜任亿级节点网络规模上的社区发现任务。但是其社区划分的效果与算法的迭代次数有关，应当根据具体需求，考虑数据规模与算法运行时间之间的关系，进而选择合适的迭代次数，获得满意的社区划分结果。

参考文献：

[1] 李建华, 汪晓锋, 吴鹏. 基于局部优化的社区发现方法研究现状[J]. 中国科学院院刊, 2015, 30(2): 238-247.
LI J H, WANG X F, WU P. Review on community detection methods based on local optimization[J]. Bulletin of Chinese Academy of Sciences, 2015, 30(2):238-247.

[2] CORNEIL D G, GOTLIEB C C. An efficient algorithm for graph isomorphism[J]. Journal of the ACM, 1970, 17(1): 51-64.

[3] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133.

[4] NEWMAN M E J. Modularity and community structure in networks[C]//The National Academy of Sciences of the United States of America. 2006: 8577-8582.

[5] ROSVALL M, BERGSTROM C T. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems[J]. Plos One, 2011, 6(4): e18209.

[6] NEWMAN M E J. Spectral methods for community detection and graph partitioning[J]. Physical Review E , 2013, 88(4): 042822.

[7] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3): 036106.

[8] 赵卓翔, 王轶彤, 田家堂, 等. 社会网络中基于标签传播的社区发现新算法[J]. 计算机研究与发展, 2011, 48(S3): 8-15.
ZHAO Z X, WANG Y T, TIAN J T, et al. A novel algorithm for community discovery in social networks based on label propagation[J]. Journal of Computer Research and Development, 2011, 48(S3): 8-15.

[9] 刘世超, 朱福喜, 甘琳. 基于标签传播概率的重叠社区发现算法[J]. 计算机学报, 2016, 39(4): 717-729.
LIU S C, ZHU F X, GAN L. A label-propagation-probability-based algorithm for overlapping community detection[J]. Chinese Journal of Computers, 2016, 39(4): 717-729.

[10] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.

[11] LIU Q, LIU C, WANG J, et al. Evolutionary link community structure discovery in dynamic weighted networks[J]. Physica A: Statistical Mechanics and its Applications, 2017, 466:370-388.

[12] FARKAS I, ÁBEL D, PALLA G, Weighted network modules [J]. New Journal of Physics, 2007, 9(6): 180.

[13] 高学东, 王立敏, 马红权, 等. 基于共享最近邻探测社团结构的算法[J]. 系统工程理论与实践, 2009, 29(10): 102-109.
GAO X D, WANG L M, MA H Q, et al. Detecting community structure based on shared nearest neighbor[J]. Systems Engineering Theory and Practice, 2009, 29(10): 102-109.

[14] 刘文远, 王佳楠, 王林. 基于局部扩张查询的重叠社区发现[J]. 小型微型计算机系统, 2015, 36(10): 2229-2234.
LIU W Y, WANG J N, WANG L. Community detection based on local expansion query[J]. Journal of Chinese Computer Systems, 2015, 36(10): 2229-2234.

[15] WICKRAMAARACHCHI C, FRINCU M, SMALL P, et al. Fast parallel algorithm for unfolding of communities in large graphs[C]//2014 IEEE High Performance Extreme Computing Conference (HPEC). 2014:1-6.

[16] 乔少杰, 郭俊, 韩楠, 等. 大规模复杂网络社区并行发现算法[J]. 计算机学报, 2017, 40(3): 687-700.
QIAO S J, GUO J, HAN N, et al. Parallel algorithm for discovering communities in large-scale complex networks[J]. Chinese Journal of Computers, 2017, 40(3): 687-700.

[17] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): P10008.

[18] STAUDT C L, MEYERHENKE H. Engineering parallel algorithms for community detection in massive networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2016, 27(1): 171-184.

[19] LU H, HALAPPANAVAR M, KALYANARAMAN A. Parallel heuristics for scalable community detection [J]. Parallel Computing, 2015, 47:19-37.

[20] AKSHAY U B. Scalable community detection using label propagation and map-reduce[R]. 2012.

[21] 从玉桓. 基于MapReduce的社区挖掘算法[D]. 上海: 上海交通大学, 2013.
CONG Y X. Community detection based on MapReduce[D]. Shanghai: Shanghai Jiao Tong University, 2013.

[22] ZHANG Q, QIU Q, GUO W, et al. A social community detection algorithm based on parallel grey label propagation[J]. Computer Networks, 2016, 107:133-143.

[23] BAE S H, HOWE B. GossipMap: a distributed community detection algorithm for billion-edge directed graphs[C]//The International Conference for High Performance Computing, Networking, Storage and Analysis. 2015: 1-12.

[24] 李春英, 汤庸, 林海, 等. 基于标签传播的可并行复杂网络重叠社区发现算法[J]. 中国科学: 信息科学, 2016, 2:212-227.
LI C Y, TANG Y, LIN H, et al. Parallel overlapping community detection algorithm in complex network based on label propagation[J]. Science China Information Sciences, 2016, 2:212-227.

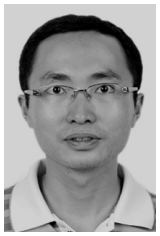
[25] PENG C, ZHANG Z, WONG K C, et al. A scalable community detection algorithm for large graphs using stochastic block models[C]//The 24th International Joint Conference on Artificial Intelligence. 2015: 2090-2096.

[26] YANG J, LESKOVEC J. Defining and evaluating network communities based on ground-truth[J]. Knowledge and Information Systems, 2012, 42(1):181-213.

[27] KWAK H, LEE C, PARK H, et al. What is Twitter, a social network or a news media?[C]//The 19th International Conference on World Wide Web. 2010: 591-600.

[28] LEUNG I X Y, HUI P, LIO P, et al. Towards real-time community detection in large networks[J]. Physical Review E, 2009, 79(6): 066107.

[作者简介]



刘强（1981-），男，江苏句容人，国防科技大学博士生，主要研究方向为社交网络分析、数据挖掘、复杂网络等。

贾焰（1960-），女，四川成都人，国防科技大学教授、博士生导师，主要研究方向为社交网络分析、信息安全等。

方滨兴（1960-），男，江西万年人，中国工程院院士，北京邮电大学教授、博士生导师，主要研究方向为社交网络分析、信息安全等。

周斌（1971-），男，江西南昌人，国防科技大学教授、博士生导师，主要研究方向为社交网络分析、信息安全等。

胡玥（1993-），女，陕西宝鸡人，国防科技大学硕士生，主要研究方向为社交网络分析。

黄九鸣（1981-），男，福建安溪人，国防科技大学讲师，主要研究方向为社交网络分析、信息安全等。