CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 44 卷第 2 期 2022 年 2 月
Vol. 44, No. 2, Feb. 2022

文章编号:1007-130X(2022)02-0306-06

一种基于 GN 算法的动态图划分方法*

罗晓霞,王　佳,罗香玉,李嘉楠
(西安科技大学计算机科学与技术学院,陕西 西安 710054)

摘　要:随着图规模的急剧增长,对动态图进行实时处理的需求日益增加。大多现有的算法针对静态图划分是有效的,直接用其处理动态图会带来较大的通信开销。针对该问题,提出一种基于 GN 算法的动态图划分方法。首先收集一段时间内加入动态图中的顶点;然后,利用 GN 算法对这些新加入的顶点进行预划分,产生若干个内部联系紧密的社区;最后,将预划分产生的社区结果插入到已经划分好的当前图中。实验从交叉边数和负载均衡度两方面将该方法与传统流式划分方法进行比较,结果表明,在公开数据集上,该方法的交叉边数降低了 13%,负载均衡度减少了 42.3%。由此可见,该方法的划分质量明显优于传统的流式划分方法。

关键词:动态图划分;GN 算法;交叉边;负载均衡度
中图分类号:TP399　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2022.02.016

A dynamic graph partitioning method based on GN algorithm

LUO Xiao-xia, WANG Jia, LUO Xiang-yu, LI Jia-nan
(College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: With the rapid growth of graph scale, the demand for real-time processing of dynamic graphs is increasing. Most of the existing algorithms are effective for static graph partitioning, but directly using them to process dynamic graphs will bring greater communication overheads. To solve this problem, a dynamic graph partitioning method based on GN algorithm is proposed. Firstly, the vertices to be inserted in the dynamic graph over a period of time are collected. Then, the GN algorithm is used to pre-partition these newly inserted vertices to generate several internally connected communities. Finally, the community results generated by the pre-partitioning are inserted into the current graph that has been partitioned. Through experiments, the proposed method is compared with the traditional streaming partitioning method in terms of crossed edges and load balance. The results show that, on the public datasets, The proposed method can reduce the number of crossed edges by 13%, and the load balance by 42.3%. It can be seen that, compared with the streaming partitioning method, the proposed method can significantly improve the quality of dynamic graph partitioning.

Key words: dynamic graph partitioning; GN algorithm; crossed edge; load balance

* **收稿日期**:2020-08-16;**修回日期**:2020-10-22
基金项目:国家自然科学基金(61702408,51634007);陕西省自然科学基础研究计划(2019JM-020)
通信地址:710054 陕西省西安市西安科技大学计算机科学与技术学院
Address: College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, Shaanxi, P. R. China

1 引言

图已经被广泛应用于医疗、食品、环保、气象和生物等各个领域[1,2]。在现实世界中,图在不断地演化,例如社交网络中[3],新成员的加入、新关系的建立、成员之间联系频率的变化,这些都会引起图的演化。并且随着图数据规模的急剧增长,单台计算机已无法对其进行处理,因此分布式图计算平台日益流行[4,5]。对动态图数据进行高效的划分处理,是提高分布式图计算效率的有效手段。

通过对图划分方法的深入研究发现,目前的一些算法主要是针对静态图划分的研究[6,7]。Hash 划分、面向 BSP(Bulk Synchronous Parallel)模型的负载均衡 Hash 图数据划分 BHP(Balanced Hash Partition)、Range 划分、基于平衡标签传播的图划分 BLP (Balanced Label propagation for Partitioning)等,这类算法适用于图结构稳定、不发生变化和不需要实时响应的静态图。当处理随时间动态演化的图时,研究人员大多采用流式划分方法。Stanton 等人[8]考虑了各种启发式方法来将图顶点分配给处理器节点,并且必须在图顶点添加到图分区系统时进行划分。Tsourakakis 等人[9]提出可扩展的流图划分框架 FENNEL,通过重新设计目标函数来考虑传统平衡图分区问题的硬约束以及切边成本。Nishimura 等人[10]将流式划分的过程变为迭代过程,图的顶点能够重新分配给处理器节点。骆融臻[11]设计并实现了一个能够可靠使用的分布式流式图划分系统,每个子分类器通过全局的共享状态进行同步,但存在较高的通信延迟。张梦琳[12]针对动态图结构,提出了一种有向性动态维护策略,通过判断图更新操作是否涉及边界顶点而给出不同的逻辑移动策略。李茜锦[13]提出一种流图分割方法,解决了有向图流分割过程中的信息丢失问题。Lü 等人[14]基于优先级的调度算法为重要顶点指定较高的优先级,以进行有效的处理,可以缩短收敛时间。以上关于动态图划分的研究,将顶点的当前邻居信息作为划分依据,并没有考虑将来一段时间内顶点邻居信息的变化。当已划分的顶点邻居信息发生变化时,需要对这些顶点进行转移,这将会带来较大的顶点转移开销,降低图划分质量。

为了解决该问题,本文提出了一种基于 GN (Girvan and Newman)算法[15]的动态图划分方法。考虑到未来一段时间内顶点邻居信息的变化,先收集一段时间内的若干个顶点邻居信息变化操作,利用 GN 算法对新加入的顶点进行预划分;然后将预划分产生的社区结果插入到当前分区中,完成动态图的划分。

2 图划分的相关理论

2.1 GN 算法

GN 算法是一种社区发现算法,本质是基于聚类的分裂思想,使用边介数作为相似度的度量方法,边介数是指图中任意 2 个顶点通过此边的最短路径的数目。GN 算法步骤如下所示:

首先计算图中所有边的边介数;然后找到边介数最大的边并将它从图中移除,计算此时的模块度;接下来重新计算图中剩下的边介数;重复上述步骤,直到图中所有的边都被删除,每个顶点单独成为一个社区为止。因为 GN 算法不能判断算法的终止位置,所以 Newman 引入了模块度的概念,用来衡量社区的划分是不是相对比较好的结果,比较每次划分之后的模块度,将模块度最大的划分结果作为最终社区划分结果。模块度计算公式如(1)所示:

$$Q=\sum_{c\in C}\left(\frac{l_c}{m}-\left(\frac{D_c}{2m}\right)^2\right) \tag{1}$$

其中,c 表示图中的一个社区;C 为所有社区的集合;m 表示图中的总边数;l_c 表示社区 c 中所有内部边的条数;D_c 表示社区 c 中所有顶点的度之和。

使用 GN 算法可以较好地发现网络中存在的社区结构,该算法对存在孤立顶点的网络、全连接社区、无权图和高内聚网络等特殊形式,均表现出良好的鲁棒性。

2.2 图划分评价指标

图划分结果主要有 2 个评价指标:负载均衡度和交叉边数[16]。其中,负载均衡度是指图数据应该尽可能均衡地被划分到多台计算机进行处理,以充分发挥分布式计算的性能优势。交叉边是指一条边的 2 个顶点被分配在不同的子图中,交叉边数会直接影响分布式计算时网络的通信开销[17]。

负载均衡度 L 是用各分区所含顶点数的方差来衡量,其计算公式如(2)所示:

$$L=\frac{(p_1-A)^2+\cdots+(p_x-A)^2+\cdots+(p_p-A)^2}{p} \tag{2}$$

其中,p 表示分区的总个数;p_x 表示第 x 个分区中的顶点个数,$1\leqslant x\leqslant p$;A 表示图中总顶点数与分

区个数的比值,即每个分区中的平均顶点数。

交叉边数是将各个子图之间的交叉边相加得到的结果,减少交叉边数可提高各分区之间的通信效率。

3 基于GN算法的动态图划分方法

3.1 基本思想

动态图的划分问题可以描述如下:假设在 t 时刻,存在一个动态图 $G_t(V_t,E_t)$,其中,V_t 和 E_t 分别表示图 G_t 的顶点集和边集,$P=\{P_{t1},P_{t2},P_{t3},\cdots,P_{tx}\}$ 表示 t 时刻的初始划分。经过 Δt 时间之后,收集给定数量 N 的图更新操作,求取新的划分 P',同时保持较好的负载均衡度和交叉边数。

本文方法的基本思想是:将收集到的 N 个图更新操作进行分类处理,对于所有的顶点插入操作,首先用GN算法进行预划分,之后将预划分结果插入当前分区中;其余的顶点删除、边插入/删除操作,分别根据收集的信息依次更新。本文方法的核心是基于GN算法,GN算法计算边介数需要找到所有最短路径,其时间复杂度为 $O(m*n)$,总时间复杂度为 $O(m^2*n)$,所以本文方法的总时间复杂度在 m 条边和 n 个顶点的图中为 $O(m^2*n)$。

3.2 相关定义

定义1 图更新操作GUOPT(Graph Update Operation):给定图 G,对该图进行的每一次更新叫做一个图更新操作,可以通过〈$type$,$value$〉的形式表示。$type\in\{1,2,3,4\}$,1表示插入顶点操作,2表示删除顶点操作,3表示插入边操作,4表示删除边操作;$value$ 表示具体更新的顶点或者边的信息。

(1) 插入顶点操作:用〈1,$value$〉表示,$value=i$,i 是图 G 中新插入的顶点。

(2) 删除顶点操作:用〈2,$value$〉表示,$value=j$,j 是图 G 中要删除的顶点。

(3) 插入边操作:用〈3,$value$〉表示,$value=(u,v)$,u 和 v 都是图 G 中的顶点,表示在顶点 u 和 v 之间插入一条边.

(4) 删除边操作:用〈4,$value$〉表示,$value=(u,v)$,u 和 v 都是图 G 中的顶点,表示删除顶点 u 和 v 之间的边。

例如,〈1,2〉表示插入顶点2;〈2,1〉表示删除顶点1;〈3,(2,3)〉表示在顶点2和顶点3之间添加一条边;〈4,(1,3)〉表示删除顶点1和顶点3之间的边。

定义2 图更新操作集 $GUOPTS$(Graph Update Operation Set):由一段时间内发生的图更新操作组成,包含多个GUOPT操作,可以表示为:$GUOPTS=\{GUOPT_1,GUOPT_2,GUOPT_3,\cdots,GUOPT_y\}$,其中 y 表示第 y 个图更新操作。

定义3 图更新操作总次数:在动态图演化过程中,出现的所有图更新操作次数的总和。

定义4 图更新频度:使用基于GN算法的动态图划分方法对整个动态图进行划分所需要的划分次数。当给定的图更新操作集大小为 N 时,更新频度 M 的计算公式如(3)所示:

$$\text{图更新频度 } M=\left\lceil\frac{\text{图更新操作总次数}}{\text{图更新操作集的大小 } N}\right\rceil \tag{3}$$

3.3 基本步骤

以给定规模的图更新操作集 $GUOPTS$ 为划分单位,收集若干个连续的图更新操作之后,做出划分决策。基于GN算法的动态图划分方法基本步骤如下所示:

步骤1 根据式(3)计算动态图划分所需要的图更新频度 M。

步骤2 收集连续的 N 个图更新操作,组成图更新操作集 $GUOPTS$。

步骤3 对于插入顶点操作,用GN算法对 $GUOPTS$ 进行处理,将新插入顶点所构成的子图预划分,产生若干个独立的社区,然后按照边的插入和删除以及顶点的删除操作进行更新:

对于插入顶点操作,可以用〈1,i〉表示,使用GN算法对新增顶点进行预划分,得到多个社区,社区内部联系紧密,社区之间联系稀疏。对于插入边操作,可以用〈3,(u,v)〉来表示,分为2种情况:当顶点 u 和顶点 v 同属于当前图或新增图时,将(u,v)插入当前图或新增图中;当2个顶点中一个属于当前图,而另一个属于新增图时,将该边记为当前图与新增图的交叉边。对于删除边操作,可以用〈4,(u,v)〉来表示,分为2种情况:当顶点 u 和顶点 v 同属于当前图或新增图时,将(u,v)从当前图或新增图中删除;当2顶点中一个属于当前图,而另一个属于新增图时,将该边从当前图与新增图的交叉边中删除。对于删除顶点操作,可以用〈2,j〉来表示,j 是图中任意顶点的编号,无论该顶点属于当前图或新增图,在对应的点集合中删除该点的信息。

步骤4 计算预划分产生的每个社区与各个

当前分区之间的交叉边数,将各社区分别插入到与之交叉边数最多的当前分区中。

将预划分产生的社区结果插入到当前分区的具体步骤如下所示:首先,将预划分产生的每个社区结果与各个当前分区之间的交叉边数置为 0;然后,从第一个社区开始循环,遍历每一条连接该社区某个顶点与当前分区某个顶点的边,每次都将对应当前分区关联的交叉边计数值加 1,直到所有社区交叉边计数结束;最后,比较每个社区与所有当前分区的交叉边计数值,找出最大值,将社区插入到最大的交叉边计数值对应的当前分区中。

步骤 5 根据更新频度 M 判断有无未处理的操作,若有,转到步骤 2;否则结束。

该方法的基本流程如图 1 所示。

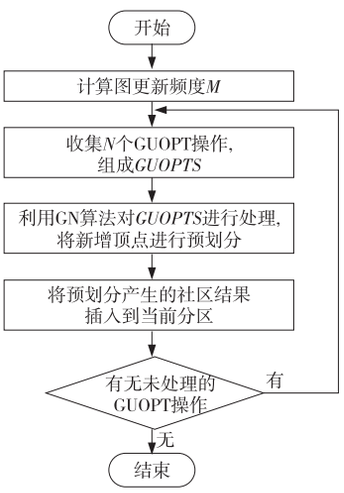


Figure 1 Basic process of dynamic graph partition

图 1 动态图划分的基本流程

4 实验与结果分析

4.1 数据集与实验环境

实验的数据集是来自斯坦福大学 SNAP(Stanford Network Analysis Project)项目组公开图数据集中的 CollegeMsg 和 Soc-sign-bitcoin-otc,具如表 1 所示。

Table 1 Data sets

表 1 数据集

数据集名称	节点数	边数
CollegeMsg	1 899	59 835
Soc-sign-bitcoin-otc	5 881	35 592

数据集 CollegeMsg 是由加州大学欧文分校的在线社交网络上发送的私人消息组成,用户可以在网络中搜索其他人,然后根据个人资料信息发起对话,边(e,f,t)表示用户 e 在 t 时刻向用户 f 发送了一条私人消息。数据集 Soc-sign-bitcoin-otc 是一个在 Bitcoin OTC 平台进行比特币交易的评价网络,边(e,f,t)表示用户 e 在 t 时刻对用户 f 进行了信用评价。由于图的演化是一个平稳的过程,针对动态图的划分,将上述 2 个数据集分别按 1∶5 的比例分为 2 部分,用少量数据作为当前图,用大量数据作为图的更新。

实验运行环境是 Windows 10,CPU 配置是 Intel(R) Core(TM) i5-4460,内存配置是 8 GB。基于 Python 3.7 实现本文提出的方法与传统的流式划分方法。

4.2 实验及结果分析

本实验分为图更新操作集大小 N 的确定与图划分质量对比 2 个阶段。

第 1 阶段:为了分析图更新操作集的大小 N 对划分质量的影响,N 分别取 1 000,2 000,3 000,4 000,5 000 和 6 000 进行实验,使用第 3 节方法完成对整个 CollegeMsg 的划分。实验结果如图 2 和图 3 所示。

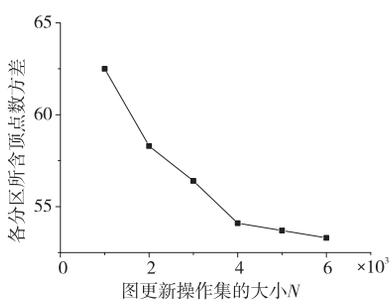


Figure 2 Load balance results

图 2 负载均衡度结果

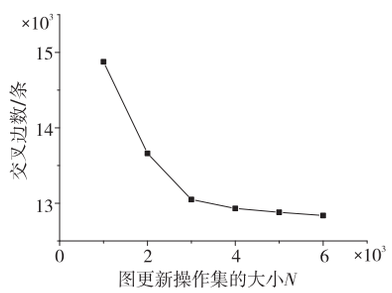


Figure 3 Crossed edges results

图 3 交叉边数结果

由图 2 和图 3 可得,随着图更新操作集大小 N 取值的增大,各分区所含顶点数方差和交叉边数的值越来越小,曲线斜率也越来越小,由此说明,图划分质量越来越好,但图更新的实时性不断地在损失。当 $N=4000$ 时,图划分质量和实时性最佳。在实际应用中,应该权衡划分质量和更新实时性 2

个方面来确定合适的图更新操作集大小 N。

第 2 阶段:分别使用传统流式划分方法和本文方法对动态图进行划分,比较划分之后的负载均衡度和交叉边数的大小。根据第 1 阶段的实验结果,给定图更新操作集的大小 $N=4000$,由式(3)计算可得,完成数据集 CollegeMsg 的划分所需要的更新频度为 13,完成数据集 Soc-sign-bitcoin-otc 的划分需要的更新频度为 8。对数据集 CollegeMsg 的划分负载均衡度对比和交叉边数对比分别如图 4 和图 5 所示,对数据集 Soc-sign-bitcoin-otc 的划分负载均衡度对比和交叉边数对比分别如图 6 和图 7 所示。

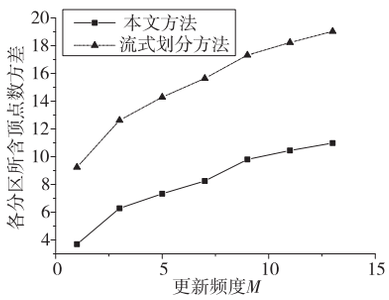


Figure 4 Comparison of load balance (CollegeMsg)

图 4 负载均衡度对比图(CollegeMsg)

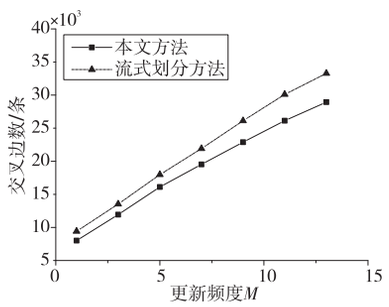


Figure 5 Comparison of crossed edges(CollegeMsg)

图 5 交叉边数对比图(CollegeMsg)

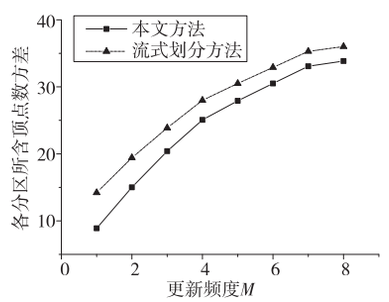


Figure 6 Comparison of load balance (Soc-sign-bitcoin-otc)

图 6 负载均衡度对比图(Soc-sign-bitcoin-otc)

由图 4 和图 6 可知,当经过相同的图更新频度 M 时,本文方法的负载均衡度曲线明显低于传统的流式划分方法的;由图 5 和图 7 可知,当经过相同的图更新频度 M 时,本文方法的交叉边数曲线

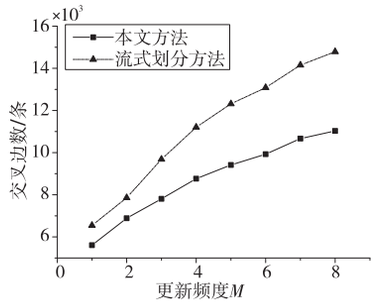


Figure 7 Comparison of crossed edges (Soc-sign-bitcoin-otc)

图 7 交叉边数对比图(Soc-sign-bitcoin-otc)

明显低于传统的流式划分方法的。由此可见,本文方法的划分结果质量明显优于流式划分方法的,各分区负载更加均衡,且产生的交叉边数更少。此外,随着图更新频度 M 的增加,本文方法的优越性更加明显,最终在完成整个图划分时,相比流式划分方法,本文方法对数据集 CollegeMsg 的划分结果中交叉边数减小了 13%,负载均衡度减少了 42.3%,本文方法对数据集 Soc-sign-bitcoin-otc 的划分结果中交叉边数减少了 25.4%,负载均衡度减少了 6%。

5 结束语

对图数据进行合理划分,是进行分布式图计算和分析的基础和前提。目前针对静态图的划分研究已经比较成熟,为了进一步提高动态图的划分质量,本文提出了基于 GN 算法的动态图划分方法,对图更新操作集内的新增图进行社区划分,再将新增图以社区为单位划分至与之联系紧密的当前分区中。实验结果表明,相较于传统流式划分方法,该方法可显著地提高动态图的划分质量。未来还需要进一步研究图更新操作集大小 N 的取值对划分结果的影响。

参考文献:

[1] Chen X, Pan L. A survey of graph cuts/graph search based medical image segmentation[J]. IEEE Reviews in Biomedical Engineering, 2018, 11: 112-124.

[2] Barsky A, Munzner T, Gardy J, et al. Cerebral: Visualizing multiple experimental conditions on a graph with biological context[J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1253-1260.

[3] Nohuddin P N E, Sunayama W, Christley R, et al. Trend mining in social networks: From trend identification to visualization[J]. Expert Systems, 2014, 31(5): 457-468.

[4] Low Y, Gonzalez J, Kyrola A, et al. Distributed GraphLab: A framework for machine learning in the cloud[J]. Proceedings of the VLDB Endowment, 2012, 5(8): 716-727.

[5] Capponi A, Fiandrino C, Kliazovich D, et al. A cost-effective distributed framework for data collection in cloud-based mobile crowd sensing architectures[J]. IEEE Transactions on Sustainable Computing, 2017, 2(1): 3-16.

[6] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49(2): 291-307.

[7] Barnard S T, Simon H D. Fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems[J]. Concurrency & Computation Practice & Experience, 1994, 6(2): 101-117.

[8] Stanton I, Kliot G. Streaming graph partitioning for large distributed graphs[C]// Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012: 1222-1230.

[9] Tsourakakis C, Gkantsidis C, Radunovic B, et al. FENNEL: Streaming graph partitioning for massive scale graphs[C]// Proc of the 7th ACM International Conference on Web Search and Data Mining, 2014: 333-342.

[10] Nishimura J, Ugander J. Restreaming graph partitioning: Simple versatile algorithms for advanced balancing[C]// Proc of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013: 1106-1114.

[11] Luo Rong-zhen. Implementation of a distributed streaming graph partition system[D]. Nanjing: Nanjing University, 2019. (in Chinese)

[12] Zhang Meng-lin. Research on partition methods on large-scale dynamic graph based on multi-level division[D]. Shenyang: Liaoning University, 2018. (in Chinese)

[13] Li Xi-jin. Streaming graph partition algorithm based on dynamic reverse mapping graph[J]. Modern Computer, 2018 (8): 89-93. (in Chinese)

[14] Lü X Q, Xiao W, Zhang Y, et al. An effective framework for asynchronous incremental graph processing[J]. Frontiers of Computer Science, 2019, 13: 539-551.

[15] Nie Xiang-lin, Zhang Yu-mei, Wu Xiao-jun, et al. A community detection algorithm based on node dependence and similar community fusion[J]. Computer Engineering & Science, 2017, 39(7): 1273-1280. (in Chinese)

[16] Xu Jin-feng, Dong Yi-hong, Wang Shi-yi, et al. Overview of large-scale graph data partitioning algorithms [J]. Telecommunications Science, 2014, 3(7): 106-112. (in Chinese)

[17] Xu Jin-feng. Large-scale dynamic adaptive graph partitioning algorithm [D]. Ningbo: Ningbo University, 2015. (in Chinese)

附中文参考文献:

[11] 骆融臻. 分布式流式图划分系统的设计与实现[D]. 南京:南京大学,2019.

[12] 张梦琳. 基于多层次划分的大规模动态图分割方法研究[D]. 沈阳:辽宁大学,2018.

[13] 李茜锦. 基于动态反向映射图的流图划分方法[J]. 现代计算机,2018(8):89-93.

[15] 聂祥林,张玉梅,吴晓军,等. 基于节点依赖度和相似社团融合的社团结构发现算法[J]. 计算机工程与科学,2017,39(7):1273-1280.

[16] 许金凤,董一鸿,王诗懿,等. 大规模图数据划分算法综述[J]. 电信科学,2014,3(7):106-112.

[17] 许金凤. 大规模动态自适应图划分算法[D]. 宁波:宁波大学,2015.

作者简介:

罗晓霞(1964 -),女,陕西西安人,教授,研究方向为大数据与云计算、人工智能与信息处理、软件工程和应用软件开发。**E-mail**: Luoxx@xust.edu.cn

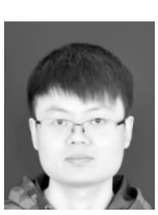
LUO Xiao-xia, born in 1964, professor, her research interests include big data & cloud computing, artificial intelligence & information processing, software engineering, and application software development.

王佳(1995 -),女,陕西延安人,硕士生,研究方向为分布式计算。**E-mail**: 1443751599@qq.com

WANG Jia, born in 1995, MS candidate, her research interest includes distributed computing.

罗香玉(1984 -),女,河北宁晋人,博士,讲师,研究方向为分布式计算。**E-mail**: 67691531@qq.com

LUO Xiang-yu, born in 1984, PhD, lecturer, her research interest includes distributed computing.



李嘉楠(1992 -),男,陕西西安人,硕士生,研究方向为分布式存储。**E-mail**: 283632680@qq.com

LI Jia-nan, born in 1992, MS candidate, his research interest includes distributed storage.