Hadoop 下并行遗传算法研究及在应急设施选址中的应用

张刚红
(兰州交通大学机电技术研究所 兰州 730070)

摘 要 随着云计算的出现,大数据的概念也随之产生。自然灾害日趋增多,要求应急设施的部署规模不断扩大,这时,如何有效进行大规模应急设施的选址成为应急管理系统的关键。因此,提出一种改进的并行遗传算法并在 Hadoop 平台上编程实现,并应用于求解应急设施选址问题的集合覆盖模型,达到求解应急设施选址的目的。试验结果表明,改进的并行遗传算法不管在获取全局最优解上还是在求解大规模应急设施选址的时效性上都优于原有算法,是一种云计算环境下有效的应急设施选址问题求解算法。

关键词 云计算;大数据;应急设施选址;Hadoop 平台;遗传算法

Research on a Parallel Genetic Algorithm in Hadoop and Application in the Site Selection of Emergency Facilities

Zhang Ganghong
(Mechanical T&LR Institute, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract With the development of cloud computing, big data emerges. The scale of first-aid facilities must be expanded because of more and more natural disasters. At the same time, it is the key of emergency management system that how large-scale site selection is finished. So a parallel genetic algorithm is presented and coded in hadoop platform. It is used to solve the set covering model of the site selection of emergency facilities to finally settle that. Testing result proves that the two aspects of the improved algorithm are better than before. One aspect is getting globally optimal solution, the other is more time efficiency. It is effective to solve the site selection of emergency facilities in cloud computing.

Key words cloud computing, big data, site selection of emergency facilities, hadoop platform, genetic algorithm

1 引言

随着云时代的来临,大数据(big data)受到了越来越多的关注。在大数据处理方面,Hadoop 是一个能对大量数据进行分布式处理的软件框架,它涵盖了云计算 3 大支撑技术中的 2 个,即并行数据处理和文件系统。其中,并行数据处理对应着通过 Hadoop 中集成的 MapReduce 模型架构;文件系统对应着 Hadoop 提供的 HDFS(Hadoop distributed file system,Hadoop 分布式文件系统)。它也是在互联网行业处理大数据的一个非常好的分布式文件系统。利用 Hadoop 挖掘大数据,利用大数据开发更大的价值,探索富有创新的空间[1]。在全国人口不断增加,自然灾害日趋增多的今天,应急设施的选址问题作为应急管理系统建立的首要任务,对应急资源的有效保障起着关键作用。应急设施选址问题是在多约束条件下,搜寻目标函数的最优解,确定设施的布局。应急设施布局以公平和强时效性作为选址标准,即在突发事件发生后,必须在最快的时间内开展有效的应急工作,因此,在应急设施的选址中,应结合应急设施的应用背景选择适合的应急设施模型,考虑应急选址问题的时效性,研究其布局问题[2,3]。随着应急设施选址规模的不断扩大,约束条件的不断增加,需要处理大规模的数据,而现有求解应急设施选址模型的算法在求解速度和求解规模上都不能满足

要求。因此,提出一种基于 Hadoop 云计算平台改进的并行遗传算法来求解应急设施选址中的集合覆盖问题,利用 Hadoop 的分布式处理技术来克服现有求解算法模型在求解速度和求解规模上的不足。最后利用测试数据来对提出的并行算法进行检验。

2 改进的并行遗传算法及应用

遗传算法的提出是基于达尔文进化论中的适者生存原理和 Mendel 遗传学说的基因遗传原理。结合这 2 种思想,就形成了遗传算法的思路,即用编码描述问题的参数,用适应度函数作为评价的依据,以编码后的群体作为进化基础,对群体中的个体位串实现选择、交叉和变异操作,并构成一个完整的迭代过程[4]。然而,传统的遗传算法不能直接在 Hadoop 上编程实现,需要对现有的遗传算法进行改进,形成可在 Hadoop 平台上容易实现的并行遗传算法,进而求解应急设施选址中的集合覆盖问题。遗传算法的具体改进方法和应急设施选址中集合覆盖问题将分别在第 2.1 节和 2.2 节中说明。

2.1 Hadoop 上改进的并行遗传算法

在 Hadoop 平台上可运行程序的模型大多采用 Google 提出的 MapReduce 模型,它特别适用于产生和处理大规模的数据集。其执行过程如图 1 所示。

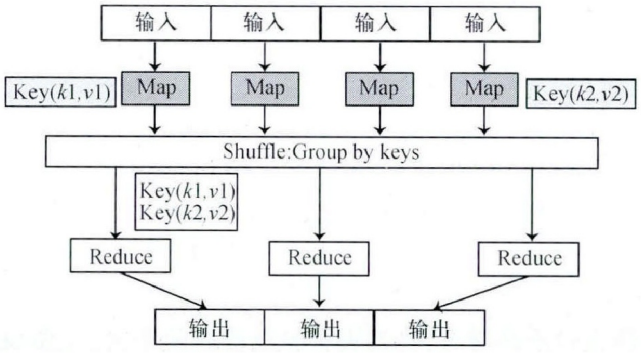

图 1 MapReduce 模型具体执行过程

从图 1 可以看出,MapReduce 有 6 个过程,可分为 2 个主要阶段。

Map 阶段:把一个较大的任务通过 MapReduce 函数分割为 M 个较小的子任务,然后配给多个 worker (被分配为执行 Map 操作的 worker)并行执行,输出处理后的中间文件。

Reduce 阶段:将 Map 阶段处理后的结果进行汇总分析处理,输出最后的结果即 R 个输出文件(R 为 Reduce 任务的个数)[5,6]。

针对遗传算法的各阶段(适应度计算、选择、交叉、变异)特点,在参考文献[7]中提出了并行遗传算法,并利用 MapReduce 进行了编程实现。针对本文求解问题的特点,提出全局并行遗传算法的思想并加入初始种群驱动函数和 Hadoop 作业迭代驱动函数,以便达到并行遗传算法更加合理且易于实现的目的。

(1)Map 函数

在 Map 函数中读取种群并计算每个个体的适应度,以<个体基因,适应度>的键值对进行输出。

Map 函数伪代码如下:

```
输入:key,value
输出:基因个体,适应度值
Map(key,value)
{
   While(存在基因个体未进行处理)
   {
      计算下一个未处理基因个体的适应度值;
      输出<当前刚处理的基因个体,适应度值>;
   }
}
```

(2)Reduce 函数

在 Reduce 函数中接受 Map 函数的输出,进行选择、交叉、变异,并将适应度值最大的个体直接进入下一代,并将下一代种群写入 HDFS 中。

Reduce 函数伪代码如下:

```
输入:key,value
输出:新的个体,NULL(空)
Reduce(key,value)
{
      根据某种策略进行选择操作;
      个体交叉;
      个体变异;
      输出<新基因个体,NULL>;
}
```

(3)驱动函数

初始种群驱动函数在 Hadoop 作业开始之前生成初始种群,而整个模型的求解需要进行多次迭代,直到满足退出条件为止,因此,Hadoop 作业迭代驱动函数负责迭代执行 Hadoop 作业,直到满足退出条件为止,其中,每次迭代 Hadoop 作业的输入是上一次作业的输出。

2.2 应急设施选址问题的集合覆盖模型

集合覆盖问题最初由 Toregas 等人于 1971 年提出,即考虑如何确定最小的设施数以及满足覆盖所有

需求点的要求。一个需求点被设施覆盖就是需求点到设施的距离要小于设施的覆盖距离，即时间限制因素的影响。集合覆盖理论主要用于消防车、医院紧急救护车等应急服务设施的选址问题中[8]。

设 $S=\{S_i|i=1,2,\cdots,m\}$ 为应急点集合，$F=\{F_j|j=1,2,\cdots,n\}$ 为设施待选点集合，w_i 为应急点 S_i 的权重，D 表示应急规定的限定距离(或 T 表示应急规定的限定时间)，d_{ij} 表示应急 S_i 点到待选点 F_j 的距离，$N_i=\{j|d_{ij}\leqslant D\}$ (或 $N_i=\{j|t_{ij}\leqslant T\}$)。假设每个设施待选点的成本不同，每个设施待选点的成本为 c_i。同时，在实际问题中，一个科学的应急限制期还应该避免系统费用的大幅增加，以便有效地保护人民及国家生命财产安全。为此，在集合覆盖问题的基础上，进行了有意义限制下应急选址问题的研究[9,10]。

设 t_{ij} 表示应急设施 F_j 到应急点 S_i 的时间，对不同的值，按照从小到大的顺利排列，记 $\{t_1,t_2,\cdots,t_k\}$，设定相应的应急限制期为 $d_k\in\{t_k,t_{k+1}\}(k=1,2,\cdots,K-1)$，定义 $A^k=a_{ij}^k$。另外，对应急点可能发生事故或灾难严重程度的不同进行考虑，需要的待选设施点数目也不同，因此，规定某一个应急点 S_i，在发生应急事件时，满足规定应急限制期内的至少 b_i 个服务设施可到达应急点 S_i 的事故现场并进行应急管理，因此，考虑建设费用集合覆盖问题的数学模型可描述为

$$\min\ z=\sum_{j=1}^{n}c_j x_j \tag{1}$$

$$\text{s.t.}\sum_{j\in N_i}x_j\geqslant b_i,\quad i=1,2,\cdots,m \tag{2}$$

$$x_j\in\{0,1\},\quad i=1,2,\cdots,m;\quad j=1,2,\cdots,n \tag{3}$$

$$y_{ij}\in\{0,1\},\quad i=1,2,\cdots,m;\quad j=1,2,\cdots,n \tag{4}$$

目标函数(1)表示设置的应急设施数最小；不等约束式(2)表示至少 b_i 个服务设施被一个设施待选点覆盖；决策变量 $x_j=1$ 表示在 j 点建立设施，否则表示不建；决策变量 $y_{ij}=1$ 表示应急点 i 被设施 j 服务，否则 $y_{ij}=0$。

2.3 求解应急设施选址中的集合覆盖模型

根据遗传算法的各项操作，对集合覆盖问题采用改进并行遗传算法的步骤如下。

(1)与启发式算法生成初始种群。

(2)计算种群中的个体适应度 f，$f=a\times f_1+b$，其中，$a=0.5\times f_{avg}/(f_{max}-f_{avg})$，$b=f_{avg}\times(f_{max}-1.5\times f_{avg})/(f_{max}-f_{avg})$，$f_1$ 为当前种群个体的适应度且 $f_1=f'_{max}-\sum_{j=1}^{n}c_j x_j$，而 f_{max}、f_{min} 和 f_{avg} 分别为 f_1 的最大值、最小值和平均值[7]。

(3)引入修补操作对种群中的不可行解进行处理。

(4)引入启发式算法思想来处理种群中的重复个体。

(5)通过跨时代精英选择选出部分个体作为交叉变异的个体。

(6)对交叉变异的个体两两配对，按照一定的交叉概率，采用自适应交叉算法进行交叉操作，交叉概率 P_c 为

$$P_c=\begin{cases}\dfrac{k_1(f_{max}-f')}{f_{max}-f_{avg}}, & f'\geqslant f_{avg}\\ k_2, & f<f_{avg}\end{cases}$$

其中，f' 表示交叉个体中的较大适应度。

(7)用自适应多位变异对交叉后的子代个体进行变异操作，变异概率 P_m 计算公式如下

$$\begin{cases}\text{round}(mu\times(me+\dfrac{f_{max}-f_1}{f_{max}-f_{min}})/1+\exp(mg\times(t-md))), & f_1>0\\ \text{round}(mu/(1+\exp(mg\times(t-md))), & f_1=0\end{cases}$$

其中，$f_1>0$ 表示可行解是要采取变异操作的个体；$f_1>0$ 表示是不可行解的个体；round 表示对整式进行去整操作；t 表示进行迭代的次数；mu 指在迭代最终达到稳定状态时的变异位数；md、me、mg 都为参数，其中，mg 的取值按照 $md>1$ 和 $md<1$ 这 2 种情况确定；$(f_{max}-f_1)/(f_{max}-f_{min})$ 表示在当前种群中，待变异个体的优劣程度，值越小表示这个个体较优，反之为较差。

(8) 在当前种群中随机找一个适应度大于 f_{avg} 的个体，用子个体替代它。

(9)采用最优保存策略。

(10) 判断结果是否满足终止条件，不满足则令 $t=t+1$，返回第二步继续进行迭代操作；满足则输出当前种群中适应度 f 值最大的个体，通过解码得出本次算法的近似最优解。

3 试验结果及分析

3.1 试验环境

本实验中采用 3 台普通台式机，其硬件配置见表1 所列。

3.2 试验数据与参数配置

试验数据以 JEBeasley 在 OR-Library[11] 网站上提供的数据为测试用例数据，选用其中 2 个用例进行试验测试改进并行算法的有效性，已知 2 个用例的最优解，其中，scpcyc06 为不带权重的集合覆盖用例数据，scpb42 为带权重的集合覆盖用例数据。算法

表 1 试验环境

机编号	硬件配置	IP 地址	在集群中的职责
机器 1	Intel-Pentium T2080(双核),内存 2 G,硬盘 160 G	192.168.1.121	NameNode/JobTracker/DataNode/TaskTracker
机器 2	Intel-Pentium T2080(双核),内存 2 G,硬盘 160 G	192.168.1.122	DataNode/TaskTracker
机器 3	Intel-Pentium T2080(双核),内存 2 G,硬盘 160 G	192.168.1.123	DataNode/TaskTracker

表 2 改进算法与陈亮[12]等人的遗传算法试验结果对比

测试用例	规模(行数×列数)	最优解	本文改进算法	陈亮等人的方法	所用时间(单位:s)	
					改进算法	陈亮等人的算法
scpcyc06	240×192	60	60	68	271.942 7	565.372 7
scpb42	200×1 000	512	512	678	624.851 4	1 247.216 3

用 Java 语言在 Hadoop 平台上实现。该试验的算法结束条件为:种群进化 80 次。如果进化结果无改进或无明显改进则停止,每个用例运行 4 次,以其中最优的解为所求解。

设置试验初始参数为:基因种群规模为 100;交叉概率 P_c 为 0.2;交叉点数 Nu 根据不同的用例设置,scpb42 时为 100,scpcyc06 时为 40;参数 k 为 1,采用自适应策略,在变异过程中,md 为 30,me 为 0.5,mg 在 $t > md$ 时为-0.05,在 $t > md$ 时为-0.5,最终稳定的变异位数 mu 设为 10。

3.3 试验结果分析

表 2 中的 2 个实例均来自参考文献[12],将改进的并行遗传算法和陈亮等人的遗传算法进行比较,改进的并行遗传算法均能得到最优值,而陈亮等人的遗传算法则不能得到。同时,表 2 和图 2 可以说明本算法在时间代价上要比陈亮等人的算法小得多,且随着求解问题规模和复杂度的不断增加,其在时间代价上的优越性更加明显。可以看出,改进的并行遗传算法更有效。

为了进一步说明对遗传算法进行改进的优越性,以 scpcyc06 为例,将改进的并行算法和基本未改进的遗传算法在进行相同进化代数所需时间进行比较,形成如图 2 所示曲线图形。从图中可以看出,两算法在开始运行时,由于需要初始化程序运行所需的环境资源,导致前几代进化所需时间的增长较快,随着程序的运行,种群进化所需时间的变化趋于线性增长。由于改进的并行遗传算法运行在 Hadoop 平台提供的计算机集群上,将计算任务划分成子任务,分配到集群中各计算机上运行,便增大了种群进化的多样性,提高了算法的收敛速度,且当进化到 59 代时,得到最优解,同时也提高了算法求解问题的速度。

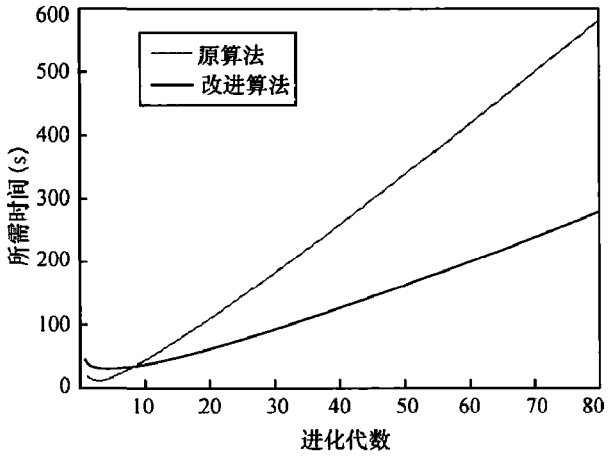

图 2 改进的并行算法与陈亮等人的算法在时间上的比较

4 结束语

应急设施选址中集合覆盖问题是 NP(non-deterministic polynomial,非确定性多项式)完全问题,传统算法很难求解精确解,而且其求解时间会随着问题规模和复杂程度的增大而急剧增加。采用改进的并行遗传算法并在 Hadoop 平台上进行编程实现来对集合覆盖问题进行求解,在求解应急设施选址中大规模集合覆盖测试用例中均能得到全局最优解。试验结果表明,改进的并行遗传算法在加快收敛速度、保证进化种群多样性、跳出局部最优解、提高算法运行速度等方面体现着其优越性。因此,改进的并行遗传算法能够有效地求解应急设施选址中的大规模集合覆盖问题。

参考文献

1 工控网.数据为王 Hadoop 与大数据处理.http://www.chinacloud.cn/show.aspx id=11626&cid=17,2012

(下转第 18 页)

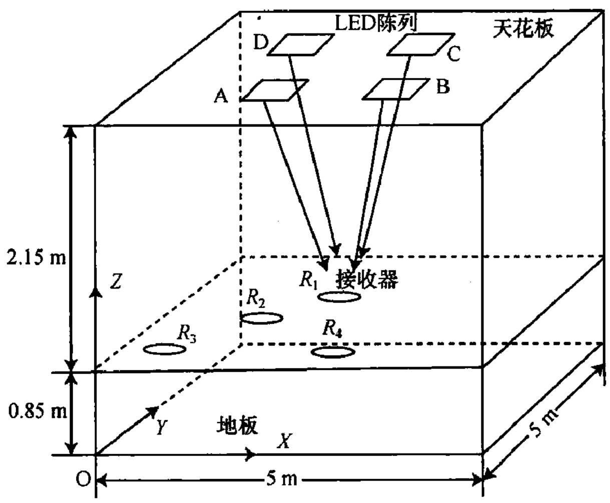

图 4 MIMO 系统组成

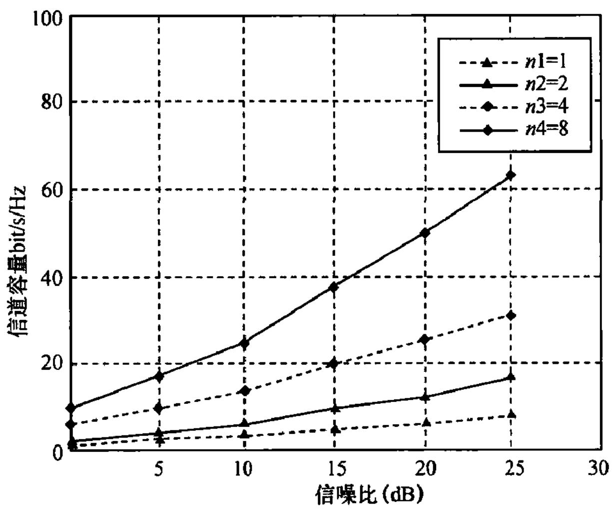

图 5 信噪比与带宽容量的关系

同时也可以增加链路的可靠性,提高光谱效率[6,7]。

根据图 3 所示建立 MIMO 系统模型，信道模型表示为 $Y(t)=RX(t)\otimes h(t)+n(t)$。其中,⊗表示卷积,R 表示光电检测器的响应率。实验仿真结果如图 5 所示，结果表明:当信噪比较低时,信道容量相差不大,随着信噪比的增加，信道容量相差很大；在一定信噪比的情况下，多个独立的信道同时传输多路数据流,增加信道容量,提高信息传输速率，通过增加信息冗余提高通信系统的可靠性。可见光通信引入 MIMO 技术,在不增加频谱资源的前提下具有更高的传输容量，还能克服通信链路因室内人员走动、家具阴影而被打断的问题。

8 结束语

可见光通信实现照明和通信 2 个功能,具有传输数据率高、保密性强、无电磁干扰、无需频谱认证等优点，与 LTE、WiMAX、Wi-Fi 和蓝牙存在互补关系,而不是所谓的替代无线传输技术。可见光通信是当前的研究热点,特别是在如何延长传输距离、自动方向对准和降低设备成本等方面。如果能成功解决这些问题,那么可见光通信将发挥巨大潜能和优势,成为无线通信领域一个新的亮点。

参考文献

1 傅倩,陈长缨,洪岳等.改善室内可见光通信系统性能的关键技术.自动化与信息工程,2010(2):4~7

2 陈特,刘璐,胡薇薇.可见光通信的研究.中兴通讯技术,2013(1)

3 丁德强,柯熙政.一种基于可见光通信的无线局域网系统设计与仿真.西安理工大学学报,2007,23(1)

4 吴华炳,谢彬,白仲亮.基于以太网的可见光通信系统.http://wenku.baidu.com/view/6d2a3df6910ef12d2af9e7b4.html,2012

5 于志刚,陈长缨,赵俊.白光 LED 照明通信系统中的分集接收技术.光通信技术,2008,32(9):52~54

6 胡国永.基于 LED 的可见光无线通信关键技术研究.暨南大学硕士论文,2009

7 卢清.基于白光 LED 室内可见光 MIMO 通信系统的研究.南京邮电大学硕士论文,2011

[作者简介] 曾庆珠,南京信息职业技术学院,教研室主任、南信院督导、副教授、工程师,主要研究方向是通信工程、通信技术。

* *

(上接第 14 页)

2 葛春景,王霞,关贤军.重大突发事件应急设施多重覆盖选址模型及算法.运筹与管理,2011,20(5):50~56

3 孙涛.青藏铁路应急救援指挥系统救援资源管理子系统的设计与实现.北京:北京交通大学出版社,2008

4 Holland J. Adaptation in natural and artificial systems. Michigan: University of Michigan Press,1975

5 李剑锋,彭舰.云计算环境下基于改进遗传算法的任务调度算法.计算机应用,2011,31(1):1~2

6 李伟卫,赵航,张阳等.基于 MapReduce 的海量数据挖掘技术研究.计算机工程与应用,2012:2~5

7 Dino Keŏ, Abdulhamit Subasi. Parallelization of genetic algorithms using Hadoop MapReduce. Southeast Europe Journal Of Soft Computing, 2012

8 余德建.应对突发事件的应急系统选址研究.南京:南京航空航天大学出版社, 2010

9 何建敏,刘春林,曹杰等.应急管理与应急系统——选址、调度与算法.北京:科学出版社, 2005

10 王小平,曹立明.遗传算法——理论、应用与软件实现.西安:西安交通大学出版社, 2002

11 Beasley J E. OR-Library. http://people.brunelac.uk/~mastjjb/jeb/info.html, 2006

12 陈亮,任世军.一种遗传算法在集合覆盖问题中的应用研究.哈尔滨商业大学学报(自然科学版),2006,22(2):67~70

[作者简介] 张刚红，兰州交通大学机电技术研究所硕士研究生,主要研究方向是云计算与数据挖掘。