

基于基因表达式编程的多数据流压缩并行算法

杨 文, 李文敬, 罗锦坤

(广西师范学院 计算机与信息工程学院, 广西 南宁 530023)

摘 要: 针对大规模数据流需要巨量存储空间, 以及串行处理速度瓶颈, 着力于解决处理后的精确度及数据压缩. 在并行平台下, 利用阈值滑动窗口技术将数据流分段送入各处理器, 并使用基因表达式编程(Gene Expression Programming, GEP)的函数发现算法实现对数据模型的函数挖掘, 提出了基于基因表达式编程的多数据流压缩并行函数替代算法 PFR-GEP (Parallel Function Replace-GEP). 在 PC 机群上的实验结果表明: 该算法有效提高了压缩比例与运算速度, 且具有线性加速比.

关键词: 并行算法; 阈值滑动窗口; 基因表达式编程; 函数发现; 多数据流; 压缩算法

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2011)08-0094-03

A Parallel Compression Algorithm for Multi-Streams Based on Gene Expression Programming

YANG Wen, LI Wen-jing, LUO Jin-kun

(School of Computer and Information Engineering, Guangxi Teachers Education University, Nanning 530023, China)

Abstract: As considering the limitation both in huge amount memory space and the speed serial processing of the large-scale flow of data, this paper aims at dealing with the accuracy and data compression after the process. In parallel platform, by taking advantage of threshold sliding window technology, the flow of data is led into each processor in sections. In addition, function mining of the data model comes true under the help of function found algorithm of the Gene Expression Programming(GEP). As mentioned above, proposing the multiple data compression algorithm based on Gene expression programming for Parallel Function Replace algorithm PFR-GEP(Parallel Function Replace-GEP). The results on PC cluster shows that the algorithm effectively improve the compression ratio and computational speed, and has a linear speedup.

Key words: parallel algorithm; threshold sliding window; GEP; function finding; multi-streams; compression algorithm

1 引言

在现代各个行业中, 如 IT、军事应用、通信等等都是以数据为基础, 对传输和存储的速度都有较高的要求, 因此数据压缩具有重要的实用价值^[1]. 在数据流压缩研究方面, 文献[2]中提出基于小波的改进算法 W-HAS-Clustering 方法并行对多数据流进行聚类分析; 文献[3]中提出了运用基因表达式编程方法来对数据进行压缩的方法. 这些研究对快速变化、

海量和实时连续的数据流压缩比或运算速度较低. 为此, 我们结合 GEP、阈值滑动窗口技术与并行的计算优势提出 PFR-GEP 算法, 提高数据流的压缩比和运行速度.

2 基本概念

2.1 基因表达式编程工作原理

(1) 初始种群

在遗传算法中通常在选择种群时, 要求群内基

收稿日期: 2011-05-10; 修回日期: 2011-06-25

基金项目: 广西自然科学基金项目(0991105); 2009 年度广西教育厅科研项目(200911MS144)

因多样化,以提高进化效率。

(2) 编码方式

对于编码的基因头部长度是可以自定,这里设定为 h ,而尾部长度 t 是对于 h 和 s 的一个函数: $t = h(s-1) + 1$,其中 s 是基因含有的函数集合中所有函数的最大操目数。

(3) 适应度函数

它代表问题处理时对环境适应能力的一种评价,也是承接问题与进化的接口。

我们使用复相关系数法: $\text{Fitness} = R^2 = 1 - \text{SSE}/\text{SST}$, $\text{SSE} = \sum_{j=1}^m (y_j - \hat{y}_j)^2$, $\text{SST} = \sum_{j=1}^n (y_j - \bar{y})^2$, 其中, $\hat{y}_j$ 表示 y_j 的估计值, $\bar{y}$ 表示 y 的平均值, SSE称为残差平方和, SST称为总离差平方和^[4]。

(4) 遗传算子

GEP的基本遗传算子有选择、变异、倒串、插串、根插串、基因变换、单点重组、两点重组和基因重组共9种^[4]。

2.2 阈值滑动窗口

定义1 滑动窗口(Sliding window)是一种流量控制技术,类似基于数组的循环操作,按照一定步长和顺序的移动与迭代中所含元素的数据集。针对阈值窗口操作的特点,在硬件实现中对循环进行一定的限制,条件如下:

条件1 循环计数的标量仅在循环过程的执行中改变;

条件2 循环的增量 $\Delta\epsilon$ 和循环边界值在循环执行过程中固定;

条件3 窗口大小由阈值来决定,与窗口位置和迭代次数无关^[5]。

设平面中 X 轴为初始线, Y 轴为阈值,如图1所示。若采用倍增法,即初始区间为 $-1 < y < 1$,下一区间为 $1 < y < 2$ (或 $-2 < y < -1$),依此类推……,其中序号为对应的分段函数。

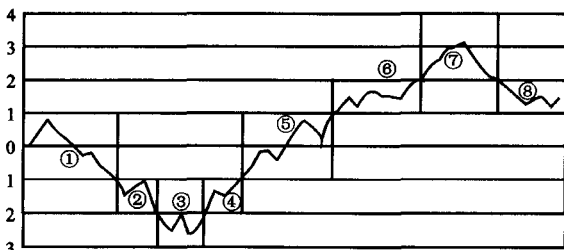


图1 阈值滑动机制

2.3 函数发现

设 x 为 m 维变量, y 为一维变量, $A = (X,$

$Y)_{(m+1) \times n}$ 为关于变量 x, y 的一组观测数据,其观测样本数为 n .其中 $A = (x_i)_{m \times n}, Y = (y_i)_{1 \times n}, i = 1, 2, \dots, m, j = 1, 2, \dots, n$.设 S 定义在 R^m 上的函数集合,则函数发现问题就是求 $f(x) = \min_{F \in S} \left\{ \sum_{i=1}^n |F(x_i) - y_i| \right\}$

3 PFR-GEP 算法过程

3.1 算法的基本思路及框架

并行函数替代算法框架如图2所示。记多数据流为 N_n ,单条数据流记为 $N_1, N_2, \dots, N_n$,依次进入主处理器;主处理器运用阈值滑动窗口机制将数据流 N_n 分成若干段 X_{np} ,在并行平台下分配到各个从处理器实现函数发现,并返回 X_{np} 对应的函数表达式 $f(x_{np})$,完成压缩存储。

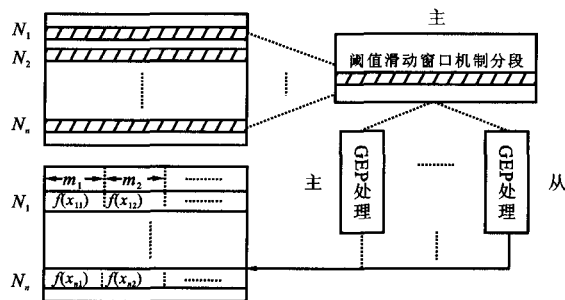


图2 并行函数替代算法模型

3.2 数据流的并行化

在给定的并行环境下,加速比取决于:(1)任务是否均衡;(2)各处理器之间的同步和通信时间;(3)任务处理速度。设有 p 个处理机,在某一时间步内,有 k 段需要计算,那么在第 i 个处理机上分配的段集为 $R_i = \{k \mid (k \bmod p) = i\} (i = 0, \dots, p-1)$,一般求解加速比有

$$S_p = T_1/T_p = T_1/(t_{\text{calc}} + t_{\text{comm}} + t_d), \quad (1)$$

式中, T_1 表示串行的运行时间, T_p 表示 p 台处理机的运行时间, t_{calc} 表示在一个时间步最慢结点机的计算时间, t_{comm} 表示处理机之间通信和同步所需时间, t_d 表示分配数据给主处理器和从处理器接收结果数据所需的时间。主要优点是每一时间步求解发现函数没有处理机之间的通信和同步开销,只有分配到处理器及返回结果所需的时间。其不足是在每一时间步,需计算的段长度不一定相同,容易使处理器的负载不均衡,从而降低加速比,若当数据段为1时退化为串行方式,对此方式,式(1)可变为

$$S_p = T_1/T_p = \sum_{j=1}^k t_j / \left(\max_{i=0, \dots, p-1} \left(\sum_{j \in R_i, j=1, \dots, k} t_j \right) + t_d \right) \quad (2)$$

式中, t_j 是当前运行的数据流中第 j 段的计算时间, 结合阈值滑动窗口技术, 使加速比迅速提高。

3.3 PFR-GEP 算法

综上述分析, 为解决速度、精确度、重现准确度等问题, 融入上述技术提炼出如下算法:

输入: 数据流 X ;

输出: 数据流段 X_n 对应的函数表达式 $f(x_{np})$.

```

1. 读入主存储数据流  $X_n$ ;
2. 根据阈值滑动窗口机制, 将数据流分成  $p$  段;
3. for ( $k = 1; k < n; k++$ ) /* 对  $n$  个数据流依次送入 */
{
    for ( $s = 1; s < p; s++$ ) /* 对  $p$  段数据流分别进行计算 */
    {
        4. 读入数据流段  $X_{np}$ , 由主处理器分配到各从处理器;
        5. 基因表达式编程的函数发现, 计算出  $f(x_{np})$ ;
    }
    6. 对替换后的完整数据流段函数进行输出;
}

```

4 实验结果与分析

4.1 实验数据及结果

实验环境: 包括 4 台 CPU2.5 GHz, RAM256 MB 同构的 PC 机群 Linux8.0 + MPI, C 编程语言。

实验一: 精确度实验. 实验参数: (1) 选用 2010 年下半年神马股份等股票, 共 30 条数据流; (2) 最大代数 2 000, 基因头长 7, 进化代数 200 参数如表 1 所示。

表 1 参数

遗传算子	概率	遗传算子	概率
变异	0.044	倒串	0.1
IS 插串	0.1	RIS 插串	0.1
单点重组	0.3	两点重组	0.3
基因重组	0.1	基因变换	0.1

函数模型中的函数集有: (1) 函数集合: +, -, *, /, ln, sqrt, cos, sin, exp; (2) 终止符集合: $x_{n-9+1}, \dots, x_n$. 限于篇幅, 仅列举了神马股份 2010 年 10 月 11 日至 2010 年 12 月 31 日数据实验结果, 如图 3 所示。

实验二: 压缩比实验. 实验参数: (1) 选取约 1 GB (1 048 000 kB) 大小的股票数据; (2) 小波变换与基于小波和耦合特征的数据源实验平台是其中的一

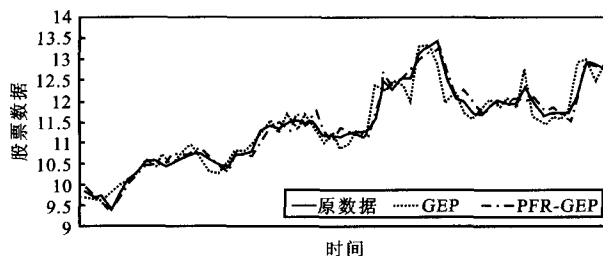


图 3 GEP 算法、PFR-GEP 算法与原数据结果对比台 PC 机, 其他同实验一. 实验结果如图 4 所示。

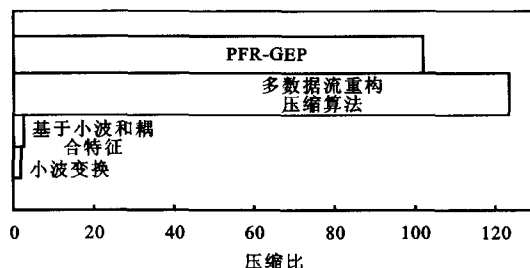


图 4 小波变换、基于小波和耦合特征、多数据流重构压缩算法及 PFR-GEP 的压缩比

4.2 结果分析

实验一分析: 从图 3 知, 使用 PFR-GEP 算法提高了压缩数据的精确度, 比 GEP 精度上误差更小, 部分具体数值可参考表 2. 此外, 算法不受数据流之间是否存在强耦合关系影响, 具有普遍性。

表 2 部分神马股份原数据及精度比较

日期	原数据	GEP	PFR-GEP	提高精度/%
2010-11-26	11.44	11.704 09	11.406 93	2.597 5
2010-11-19	11.26	10.970 18	11.259 16	2.566 4
2010-11-16	11.67	12.402 21	11.634 56	6.578 0
2010-11-15	12.74	12.241 80	12.527 98	2.246 4
2010-11-10	12.80	11.952 54	12.557 65	4.727 4

实验二分析: PFR-GEP 算法压缩后约为 10 200 kB. 因实施了分段机制, 所以 PFR-GEP 算法的压缩比例比小波变换 (约 916 800 kB) 高 100~130 倍; 比基于小波和耦合特征 (约 349 700 kB) 高 30~60 倍; 多数据流的重构算法 (约 8 500 kB), 压缩比例接近 $1/N$, N 表示数据流条数, 而 PFR-GEP 算法的压缩比例为 k/N (其中 k 为数据流段数), 故压缩比例较接近, 但在速度与精度上都有大幅度的提高。

5 结束语

本文在利用 GEP 的函数发现以及并行的优势, 对多数据流进行并行处理以达到压缩目的, PFR-GEP 算法提高了数据处理的速度及精度且在压缩

(下转第 101 页)

T/pT ,表示用户停留时长占整个节目时长的比例, pT 表示一个完整节目的时间.如图8所示为 U 的分布图,可以看到,三个节目的测量结果都相似,一部分用户在观看一段时间后就离开了该节目,另一部分用户则会在观看完节目后离开.

综上,点播系统中存在一个明显的用户行为特征:跳跃观看者数量远远大于平滑观看者数量.当用户观看点播节目时会迅速的浏览一些节目片段直到发现他们感兴趣的节目,然后就会长时间的连续观看,如果没有感兴趣的节目就会退出系统.

5 结束语

本文通过对PPTV点播系统进行协议分析,利用协议分析和逆向工程的主动测量方法,获悉了它的工作原理,并在此基础上设计实现了PPTV点播系统分布式爬虫系统,利用该爬虫完成了对PPTV点播系统覆盖网的主动测量.测量结果发现:(1)PPTV点播系统用户数量变化规律与用户的作息时间关联明显;(2)在点播系统中,跳跃观看者数量远远大于平滑观看者数量,用户会快速的浏览一些节目直到发现感兴趣的节目.本文的研究方法为P2P点播系统的测量与监控提供了参考,研究结果为P2P点播系统的监控及优化提供了依据.

参考文献:

- [1] Hei X, Liang C, Liang J, et al. A measurement study of a large-scale P2P IPTV system [J]. IEEE Transactions

on Multimedia, 2007, 9(8):1672-1687.

- [2] 李代玲. 基于网络测量的PPStream网络电视系统研究[D]. 北京:北京交通大学,2008.
- [3] 樊鹏翼. P2P流媒体监管关键技术研究[D]. 长沙:国防科学技术大学,2008.
- [4] 时金桥,张志斌,郭莉,等. 大规模分布式互联网应用的测量[J]. 信息技术快报, 2008, 6(6):1-14.
- [5] The WireShark Homepage [EB/OL]. [2010-09-30]. <http://www.WireShark.org/>.
- [6] 冯健. P2P流媒体关键技术研究[J]. 微电子学与计算机, 2009, 26(8):49-54.
- [7] 王海舟,陈兴蜀,王文贤. PPLive网络电视系统的测量研究[J]. 计算机应用, 2009(7):1988-1991.
- [8] Long V, Guta I, Liang J, et al. Measurement and modeling of a large-scale overlay for multimedia streaming [EB/OL]. [2009-06-07]. <http://dprg.cs.uiuc.edu/docs/pplive/qshine07-4.pdf>.
- [9] 姜志宏,王晖,樊鹏翼,等. 一个P2P IPTV多协议爬行器—TVCrawler[J]. 计算机应用, 2010, 30(3):58-60.

作者简介:

- 胡鑫 男,(1988-),硕士研究生. 研究方向为P2P网络.
- 陈兴蜀 女,(1968-),博士,教授,博士生导师. 研究方向为信息安全、计算机网络.
- 王海舟 男,(1986-),博士研究生. 研究方向为信息安全、P2P网络.
- 刘磊 男,(1983-),硕士研究生. 研究方向为信息安全、P2P网络.

(上接第96页)

比例上比较理想.(1)阈值滑动机制中阈值的划定(2)如何提高压缩比例及在其他领域的应用是需要进一步研究的工作.

参考文献:

- [1] Chen Yu, Tang Changjie, Li Rui, et al. Reduced-GEP: Improving gene expression programming by gene reduction[C]// Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2010 2nd International Conference. China: Hangzhou, 2010:176-179.
- [2] 陈华辉,施伯乐,钱江波,等. 基于小波概要的并行数据流聚类[J]. 软件学报, 2010, 21(4):644-658.
- [3] 丁超,元昌安,李桂来,等. 基于GEP的多数据流的压缩算法[J]. 计算机研究与发展, 2008, 45(S0):191-195.

- [4] 元昌安,彭昱忠,覃晓,等. 基因表达式编程算法原理与应用[M]. 北京:科学出版社,2010.
- [5] 刘陶刚,赵荣彩,姚远,等. 分块存储的滑动窗口数据重用技术[J]. 计算机应用, 2010, 30(5):1371-1375.
- [6] 王元元,曾建潮,谭瑛. 基于并行计算模型的并行微粒群算法的性能分析[J]. 微电子学与计算机, 2007, 24(9):25-28.

作者简介:

- 杨文 男,(1987-),硕士研究生. 研究方向为并行计算和人工智能.
- 李文敬 男,(1964-),教授. 研究方向为并行算法和演化算法.
- 罗锦坤 男,(1985-),硕士研究生. 研究方向为数据挖掘与智能计算.