第 34 卷 第 12 期 计 算 机 工 程 2008 年 6 月
Vol.34 No.12 Computer Engineering June 2008
·博士论文· 文章编号：1000—3428(2008)12—0019—03 文献标识码：A 中图分类号：TN911.72

基于 MPI 并行计算的信号稀疏分解

刘 浩[1]，杨 辉[2]，尹忠科[1]，王建英[1]
(1. 西南交通大学信息科学与技术学院，成都 610031；2. 摩托罗拉中国软件中心，成都 611731)

摘 要：在研究信号稀疏分解理论及其最常用的匹配追踪算法的基础上，针对 MP 算法存在的计算量过大的问题，提出一种基于并行计算系统实现信号稀疏分解的方法。该方法利用 8 台微机，采用 MPI 消息传递机制，以 100 M 高速以太网作为互联网络，构建了一套 Beowulf 并行计算系统，在此系统上通过编制并行程序来实现 MP 算法。实际测试表明这种方法具有很高的并行计算效率，分解时间从单机 75 min 左右下降到 8 机并行 11 min 左右，大大提高了信号稀疏分解的速度。

关键词：稀疏分解；匹配追踪；并行计算；MPI 消息传递

Signal Sparse Decomposition Based on MPI Parallel Computing

LIU Hao[1], YANG Hui[2], YIN Zhong-ke[1], WANG Jian-ying[1]
(1. School of Information Science & Tech., Southwest Jiaotong University, Chengdu 610031; 2. Motorola(China) Electronics Ltd., Chengdu 611731)

【Abstract】 After studying Matching Pursuit(MP) algorithm of signal sparse decomposition, this paper proposes a new approach to improve the speed of MP algorithm, and it describes how to build a Beowulf parallel computing system with 8 PCs. Its parallel computation is implemented by Message-Passing-Interface(MPI), and a 100Mb/s high speed Ethernet network interconnects all PCs. Test is made using parallel computing program to measure the parallel efficiency of the system, results show that this parallel can reduce the MP algorithm computing time-cost from 75 minutes with a PC to 11 minutes with 8 PCs.

【Key words】 sparse decomposition; Matching Pursuit(MP); parallel computing; MPI message passing

在信号处理研究中，传统的信号分解是将信号分解在一组完备的正交基上，如傅立叶变换、正交小波变换等。随着信号处理理论的发展，近年来信号的非正交分解引起了学者越来越多的重视。Mallat 和Zhang于 1993 年提出了信号在过完备库上分解的思想[1]：将信号分解在一组过完备的非正交基上，由分解的结果，可以得到信号的一个非常简洁的表达，即稀疏表示(sparse representation)，此过程称为信号的稀疏分解(sparse decomposition)。由于信号的稀疏表示的优良特性，信号稀疏表示已经被应用到信号处理的许多方面，如信号去噪[1]、编码和识别[2]。

1 基于匹配跟踪算法(MP)的信号稀疏分解

按照稀疏分解理论，为了得到信号的稀疏表示，必须使用非正交基(称为原子)。由这些原子组成的集合是过完备的，被称为过完备库，信号在过完备库上的分解结果一定是稀疏的[1]。

设 $D=\{g_\gamma\}_{\gamma\in\Gamma}$ 为用于进行信号稀疏分解的过完备库，g_γ 为由参数组 γ 定义的原子。原子 g_γ 的长度与信号本身长度相同，但作了归一化处理，即 $\|g_\gamma\|=1$，Γ 为参数组 γ 的集合。由库的过完备性可知，参数组 γ 的个数应远大于信号的长度。

信号稀疏分解的关键问题就是从信号集 $D=\{g_\gamma\}_{\gamma\in\Gamma}$ 中选择尽量少的基信号，通过其线性组合最佳地逼近待分解信号 f，这其实是一个组合优化问题。即

$$\min\left\|f-\sum_{k=0}c_k g_k\right\| \tag{1}$$

其中，c_k 为分解系数；g_k 为第 k 个基信号。

通常信号集是一无穷集，因此在处理中必须利用优化的办法选择最少的基信号来重构原信号。文献[1]在投影跟踪算法[3]的基础上提出一种自适应递归投影算法——匹配跟踪(Matching Pursuit, MP)算法，有效地实现了在过完备原子库中的搜索。

MP 方法分解信号过程如下：

设待分解信号 $f\in H$，信号长度为 N，H 为 Hilbert 空间。首先从过完备库中选出与 f 最大投影匹配的原子 g_{γ_0}，即满足以下最大内积条件：

$$\left|\langle f,g_{\gamma_0}\rangle\right|=\sup_{\gamma\in\Gamma}\left|\langle f,g_\gamma\rangle\right| \tag{2}$$

信号因此可以分解为在最佳原子 g_{γ_0} 上的投影分量和残差 2 部分，即

$$f=\langle f,g_{\gamma_0}\rangle g_{\gamma_0}+R^1 f \tag{3}$$

其中，$R^1 f$ 是用最佳原子对原信号进行最佳匹配后的残差，对最佳匹配后的残差可不断进行与上面同样的分解过程，即

$$R^k f=\langle R^k f,g_{\gamma_k}\rangle g_{\gamma_k}+R^{k+1} f \tag{4}$$

其中，g_{γ_k} 满足

$$\left|\langle R^k f,g_{\gamma_k}\rangle\right|=\sup_{\gamma\in\Gamma}\left|\langle R^k f,g_\gamma\rangle\right| \tag{5}$$

由式(3)和式(4)可知，经过 n 步分解后，信号被分解为

$$f=\sum_{k=0}^{n-1}\langle R^k f,g_{\gamma_k}\rangle g_{\gamma_k}+R^n f \tag{6}$$

基金项目：国家自然科学基金资助项目(60602043)；四川省应用基础研究基金资助项目(2006J13-114, 04JY029-05)

作者简介：刘 浩(1969－)，男，博士研究生，主研方向：信号处理；杨 辉，高级工程师；尹忠科，教授；王建英，副教授

收稿日期：2007-07-02 **E-mail：**hliu@home.swjtu.edu.cn

其中，$R^n f$ 为原信号分解后的残差信号。由于每一步分解中，所选取的最佳原子满足式(5)，因此分解的残余 $R^n f$ 随着分解的进行，迅速地减小，$\|R^n f\|$ 随 n 的增大而指数衰减为 0。从而信号可以分解为

$$f = \sum_{k=0}^{\infty} \langle R^k f, g_{\gamma_k} \rangle g_{\gamma_k} \tag{7}$$

实际上，由于 $\|R^n f\|$ 的衰减特性，一般而论，用少数的原子(与信号长度相比较而言)就可表示信号的主要成分，即

$$f \approx \sum_{k=0}^{n-1} \langle R^k f, g_{\gamma_k} \rangle g_{\gamma_k} \tag{8}$$

由上述可知，MP 算法首先需要完成原始(或残差)信号在过完备库中的每一个原子上的投影计算，其实质就是一个在高维(N 维)空间进行的内积计算 $\langle R^k f, g_\gamma \rangle$，然后再逐个进行比较，找出最佳原子。在常规算法中，由于采用的是传统的串行程序模式，只能按式(2)~式(8)的顺序逐个地在原子库中“计算-比较-计算……”。如果原子库较大，计算量将非常大，这是稀疏分解理论在计算机工程实现上进展缓慢的重要原因。过完备原子库中快速搜索和计算算法已成为信号稀疏分解研究中的热点。

针对 MP 算法在过完备原子库中搜索的特点，本文提出了基于 MPI 支持环境的 MP 并行处理算法来实现信号的稀疏分解。

2 基于并行算法的信号稀疏分解

2.1 Beowulf 并行计算机系统

随着科学技术的飞速发展，越来越多的大型科学与工程计算问题，如气象预报、空气动力学计算等，对计算机的计算速度提出了非常高的要求，高性能计算已在科学研究中起到了重要的作用。在高性能计算方法中，巨型机、大型机由于价格昂贵，不能普遍应用，而将许多普通的微机构建成“计算集群”(computing cluster)，通过并行计算方式，以较低的成本获得强大的计算能力，成为计算机系统发展的一个重要方向[4]。一般将这种用一组微机通过以太网连接起来的并行计算集群称为Beowulf。图 1 是Beowulf并行计算机系统的原理图。

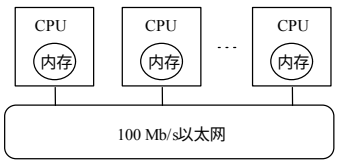

图 1 Beowulf 并行计算机系统原理

图 1 中所示并行计算系统采用的是分布式内存方式，各个处理单元都拥有自己独立的局部存储器。由于不存在公共可用的存储单元，因此各个处理器之间通过消息传递来交换信息协调和控制各个处理器的执行。一个物理问题并行求解的最终目的是将该问题映射到并行机系统上，这一物理上的映射是通过不同层次上的抽象映射来实现的[5]，如图 2 所示。

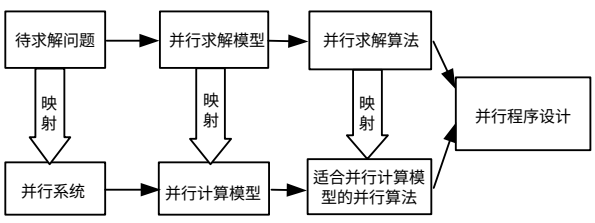

图 2 问题的并行求解流程

在图 2 中，从第 1 步(待求解问题)到第 2 步(并行求解问题)是对问题的抽象描述；从第 2 步(并行求解问题)到第 3 步(并行求解算法)是对问题的精确描述。

并行计算系统的核心和精髓是各个并行执行的节点(或者说各处理器)之间通过传递消息来交换信息、协调步伐。消息传递机制为编程者提供了更灵活的控制手段和表达并行的方法，能大大提高并行程序的执行效率。对于一般Beowulf系统来说，采用的是MPI消息传递机制[6]。

2.2 MPI 消息传递标准

Message Pass Interface(MPI)是并行计算机的消息传递接口标准。开发MPI的目的是为了提高并行程序的可移植性和易用性[5]。MPI是一个消息传递型并行通信的程序设计规范，是一个消息传递库。其库函数可以调用C, C++, Fortran 等编程语言，消息传递的并行编程主要是通过调用消息传递库MPI函数来进行的，它实现了处理器之间的数据交换功能。MPI支持多种操作系统平台。MPI的特点是免费和源代码开放，最重要的实现方法是通过MPICH软件实现。

2.3 基于 MPI 的信号稀疏分解并行计算

MPI 是基于消息传递模型的，在这种模型中，把任务分成若干部分(进程)，让每个处理器(节点)独自运行不同或者相同的部分(进程)。驻留在不同处理器(节点)上的进程可以通过网络传递消息来互相通信，从而达到并行计算的效果，减少计算的时间。用 MP 算法实现在过完备空间的搜索和计算时，由于在原子上的投影计算实质是在 N 维空间进行的内积计算，因此当原子数量较大时，运算量将非常大。对于此问题，本文提出了“基于消息传递的主从式并行算法”方案。其要点是：将原子库里的原子分为 n 组(n 为并行系统中处理器个数)，每组归属于驻留在每个不同处理器(节点)上的从进程管理，待分解信号在每组原子库里进行搜索和计算(n 个从进程同时进行)，选出内积最大者通过消息传递机制送给主进程，在主进程里再把每个从进程传来的原子用 MP 算法进行搜索和计算，完成一次迭代。整个基于并行计算的 MP 分解算法流程如图 3 所示。

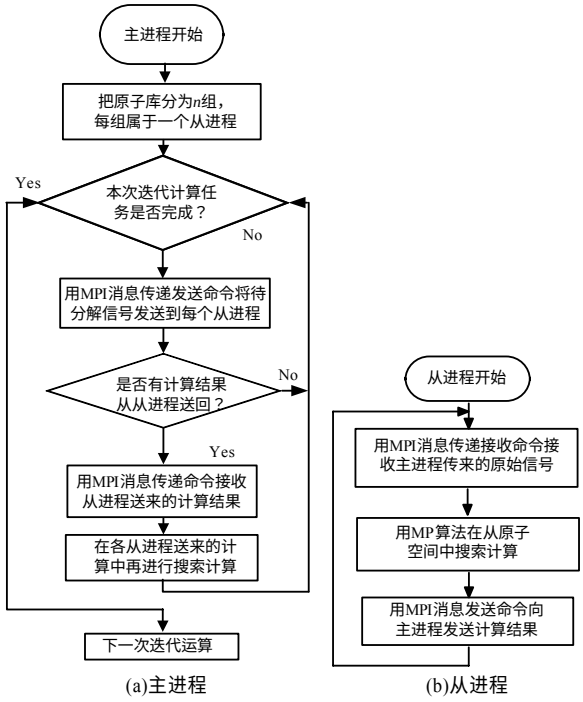

图 3 主从式并行算法实现信号稀疏分解流程

3 实际测试及结论

在实际测试中，用 8 台联想 D3500A 微机构建成一个 Beowulf 并行计算系统。每台微机的配置为 Athlon64 3600+ 处理器，512 MB DDR 内存，每个节点上运行 Windows XP 操作系统，采用 MPICH1.2 作为并行计算的支撑环境，用 100 Mb/s 高速以太网作为网络互联。

实际测试中选用Gabor原子来构成过完备原子库[1]，Gabor原子由一个经过调制的高斯窗函数组成：

$$g_\gamma(t)=\frac{1}{\sqrt{s}}g(\frac{t-u}{s})\cos(vt+w) \tag{9}$$

其中，$g(t)=\mathrm{e}^{-\pi t^2}$ 是高斯窗函数；$\gamma=(s,u,v,w)$ 是时频参数；s 是伸缩因子(尺度因子)；v 是原子的频率；w 是原子的相位。参数 s , v 和 w 定义了一个原子的形状，而参数 u 定义了一个原子的中心位置。图 4 就是一个原子的示意图。

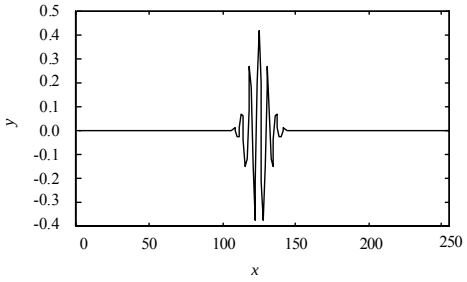

图 4　一个原子的示意图

从文献[1]可以推导出过完备原子库 D 中原子的个数 L_D，其计算公式如下：

$$L_D=52(N\mathrm{lb}N+N-1) \tag{10}$$

设待分解信号的长度 $N=256$，通过计算可知，L_D=119 756，即信号稀疏分解使用的过完备原子库中原子个数为 119 756 个。由第 2 节的分析可知，计算量将十分巨大，因此本文采用并行计算方法来实现信号稀疏分解，并和单机上一般稀疏分解方法进行比较。在每一次用 MP 算法进行信号稀疏分解时，都是在分解迭代次数定为 30 次的同一条件下比较运算时间，因为实验证明 30 次迭代可以完全分解和重构信号。

结果如图 5 所示，并行计算使整个信号稀疏分解耗时由单机 75 min 左右下降到 8 机并行 11 min 左右。

并行效率如图 6 所示，当处理器个数增加时，并行效率有所下降。

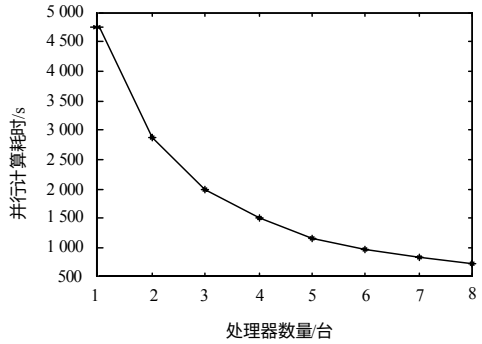

图 5　并行计算耗时

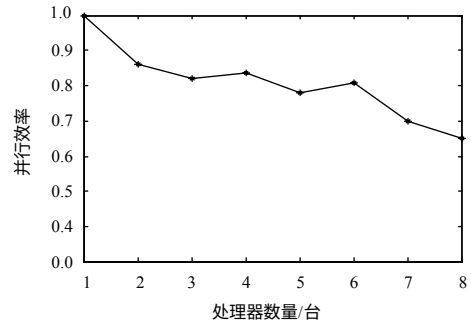

图 6　并行计算效率

从图 5、图 6 可看出，Beowulf 并行计算系统在实际性能测试中表现出了很好的计算效率，大大减少了信号稀疏分解的计算量。实际测试证明该系统的构建是成功的。

参考文献

[1] Mallat S, Zhang Z. Matching Pursuit with Time-frequency Dictionaries[J]. IEEE Trans. on Signal Processing, 1993, 41(12): 3397-3415.

[2] Arthur P L, Philipos C L. Voiced/Unvoiced Speech Discrimination in Noise Using Gabor Atomic Decomposition[C]//Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing. Hong Kong, China: [s. n.], 2003: 820-828.

[3] Huber P J. Project Pursuit[J]. The Annals of Statistics, 1985, 13(2): 435-475.

[4] 黎康保, 陶文正, 许丽华. 用 PC 机群组构并行超级计算机[J]. 计算机工程, 2000, 26(9): 1-3.

[5] 都志辉. 高性能计算并行编程技术——并行程序设计[M]. 北京: 清华大学出版社, 2001.

(上接第 18 页)

表 2　以“药病”关系建立的语义场试验结果

基本概念	相关概念
当归	月经不调、经闭腹痛、症瘕结聚、崩、漏、头痛、眩晕、痿痹、肠燥便难、痈疽疮疡、跌扑损伤
白芍	月经不调、经行腹痛、崩漏
茯苓	痰饮咳逆
香附	疟疾、脱肛、月经不调、子宫下垂

5 结束语

本文从中医理论出发，利用本体知识获取技术得到蕴含于中医临床用药诊治过程中的隐性知识及规律，从构建医案的本体知识库着手，对语义网络进行定量的深度分析，提出了概念密度的计算方法，为进一步构建中医领域的本体系统、挖掘整理中医临证经验与学术思想及建立基于信息检索技术的中医知识库打下了基础。

参考文献

[1] 陆汝铃. 人工智能[M]. 北京: 科学出版社, 2000.

[2] 肖位枢. 图论及其算法[M]. 北京: 航空工业出版社, 2005.

[3] 宋　炜, 张　铭. 语义网简明教程[M]. 北京: 高等教育出版社, 2004.

[4] Natalya F N, Deborah L M. Ontology Development 101: A Guide to Creating Your First Ontology[D]. Stanford, USA: Stanford University, 2000.

[5] 董振东, 董　强. 知网[EB/OL]. (2008-04-11). http://www.keenage.com.

[6] 王大亮, 孙建淘. 基于 HowNet 构造语义场的方法[J]. 清华大学学报: 自然科学版, 2005, 45(1): 77-80.