基于 MPI 和 OpenMP 混合编程的非负矩阵分解并行算法

唐 兵[1] Laurent BOBELIN[2] 贺海武[2]

(湖南科技大学计算机科学与工程学院 湘潭 411201)[1] (中国科学院计算机网络信息中心 北京 100190)[2]

摘 要 非负矩阵分解(NMF)作为一种数据降维和特征提取的有效工具,已经在文本聚类、推荐系统等多个领域得到应用,但是其计算过程比较复杂。对此,提出一种基于 MPI+OpenMP 的混合层次化并行 NMF 方法,其充分利用基于 MPI 的消息传递模型和基于 OpenMP 的共享存储模型各自的优势,并基于多核节点集群进行测试。实验结果表明,所设计的并行 NMF 算法达到了较高的加速比,能有效处理高阶矩阵的非负分解,极大地提高了计算的效率。

关键词 非负矩阵分解,并行算法,MPI,OpenMPI,可扩展

中图法分类号 TP319 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.03.013

Parallel Algorithm of Nonnegative Matrix Factorization Based on Hybrid MPI and OpenMP Programming Model

TANG Bing[1] Laurent BOBELIN[2] HE Hai-wu[2]

(School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China)[1]

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract Nonnegative matrix factorization (NMF) has been introduced as an efficient way to reduce the complexity of data and extracting character, and it has also been applied to various fields, such as recommendations and text clustering. However, the computation process of NMF is quite complex. In order to solve this problem, a hybrid parallel hierarchical NMF algorithm based on OpenMP and MPI was presented in this paper, which makes full use of the advantages of both MPI-based message passing model and OpenMP-based shared storage model. The new algorithm is evaluated in a multi-core cluster environment, and experimental results demonstrate that it can achieve a high speed-up, and can be used to deal with large-scale NMF with a high efficiency.

Keywords Nonnegative matrix factorization, Parallel algorithm, MPI, OpenMPI, Scalability

1 引言

非负矩阵分解(Nonnegative Matrix Factorization, NMF)是对所有元素均为非负数的矩阵进行分解,得到两个低秩矩阵,可以实现对数据的降维和特征提取[1]。与 PCA(主成分分析)、ICA(独立成分分析)、SVD(奇异值分解)、VQ(矢量量化)等矩阵分解不同,NMF 克服了传统矩阵分解的很多问题,其结果具有十分直观的语义解释。NMF 已成为了机器学习和数据挖据中一种重要的数学方法,被广泛应用于特征提取、图像分析、推荐系统[2]、模式识别、数据聚类[3]、主题模型[4]、信号分析[5]、基因数据分析[6]等众多领域。

NMF 存在的主要问题是原始矩阵的维度通常比较大,并且计算的复杂度较高。NMF 的并行化算法也逐渐受到重视,其虽然能够在一定程度上提高计算的效率,但是一个好的并行算法应该要与机器硬件体系结构相匹配,要有较强的可扩展性,即应该能有效地利用增加的处理器资源的能力。本文提出一种基于 MPI+OpenMP 的 NMF 并行处理方法,主进程分派矩阵块,然后进行迭代更新。在各计算进程内部,采用 OpenMP 线程并行。经实验验证,这种 MPI 位于顶层、OpenMP 位于底层的层次化并行方法能够提高 NMF 计算的速度,取得较好的效果。

2 MPI+OpenMP 混合并行编程模型

并行编程模型一直是并行计算研究领域中的重点内容,目前两种最重要的并行编程模型为共享存储编程模型和消息传递编程模型。共享存储模型具有单地址空间、编程容易、可移植性差等特点,其实现有 OpenMP 和 Pthreads 等。消息传递编程模型具有多地址空间、编程困难、可移植性好等特点,其实现有 MPI 和 PVM 等。

MPI 是集群计算中应用较为广泛的编程平台,但是单一的 MPI 消息传递并行编程模型在多处理器节点集群上的并行效果并不理想。MPI+OpenMP 这种层次化的并行模型能够结合分布式内存结构和共享内存结构的优势,提供节点间和节点内的两级并行。它结合了进程级的粗粒度并行(例如区域分解)和线程级的细粒度并行(如循环并行),在多数情况下,其执行效率高于纯 MPI 或 OpenMP 程序。这种 MPI 位

到稿日期:2015-12-28 返修日期:2016-03-23 本文受中科院国际人才计划 CAS PIFI(2016VTB028),湖南省自然科学基金(2015JJ3071)资助。

唐 兵(1982-),男,博士,讲师,主要研究方向为并行与分布式计算、云计算等,E-mail:btang@hnust.edu.cn;贺海武(1977-),男,博士,研究员,主要研究方向为并行与分布式计算、云计算等,E-mail:haiwuhe@cstnet.cn(通信作者)。

于顶层、OpenMP 位于底层的混合编程模型很好地映射了多处理器节点集群体系结构[7-9]。

3 非负矩阵分解

3.1 NMF 算法介绍

NMF 的思想自从 D Lee 和 H Seung 于 1999 年提出开始，就迅速得到了人们的重视[1]。其基本思想可描述为：任意给定一个非负矩阵 $V=(v_{ij})_{n\times m}=[v_1,v_2,\cdots,v_m]$，寻找一个非负矩阵 $W\in R^{n\times r}$ 和一个非负矩阵 $H\in R^{r\times m}$，满足 $V\approx WH$。

NMF 是一个 NP 问题，可以转化成优化问题并通过迭代方法求解 W 和 H，只能使分解误差尽可能小，定义适当的目标函数，通过最小化目标函数来找到 V 尽可能精确的分解。以欧氏距离为例：

$$E(V\|WH)=\|V-WH\|_F^2=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(v_{ij}-\sum_{k=1}^{r}w_{ik}h_{kj})^2 \tag{1}$$

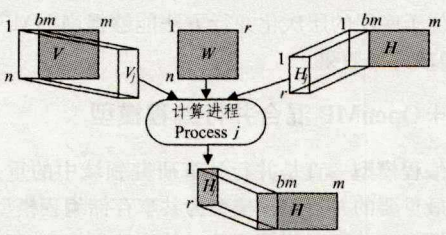
通过下列迭代可以得到上述最优化问题的局部最优解：

$$h_{ij}=h_{ij}\frac{(W^TV)_{ij}}{(W^TWH)_{ij}},w_{ij}=w_{ij}\frac{(VH^T)_{ij}}{(WHH^T)_{ij}} \tag{2}$$

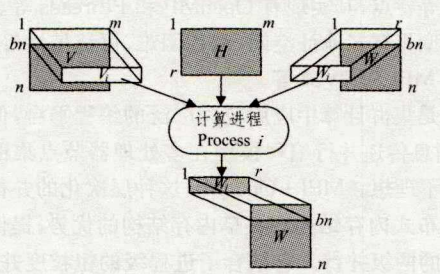
本文主要采用欧氏距离作为目标函数，采用式(2)中的迭代原理来描述 NMF 并行算法。

3.2 并行 NMF 的基本原理

通过对式(2)进行分析可得并行 NMF 迭代计算的基本原理图(见图 2)。矩阵运算采取分块的方式进行，其中 bm 的大小可根据硬件的配置进行调整。在初始化时，产生一个初始的解 W 和 H。如图 1(a)所示，矩阵 W 的大小为 $n\times r$，矩阵块 V_j 的大小为 $n\times bm$，矩阵块 H_j 的大小为 $r\times bm$，最终可得到更新后的矩阵块 H_j，通过拼接可得到新的矩阵 H。如图 1(b)所示，新的矩阵 H 被用于计算新的矩阵块 W_i，以此类推，采用 H 矩阵和 W 矩阵交替迭代的方式进行。



(a)矩阵 H 的迭代



(b)矩阵 W 的迭代

图 1 基于分块的 NMF 迭代的基本原理图

由此分析可知，V 矩阵相当于一个只读变量，需各进程间共享读取即可，但是随着迭代的进行，W 矩阵和 H 矩阵需要在各进程间进行同步。

4 基于 MPI 和 OpenMP 的混合并行 NMF 算法

如式(2)所示，单机环境下，串行 NMF 算法主要采用交替更新 W 和 H 的方法进行多次迭代以得到分解的结果。算法要求输入原始矩阵 $V_{n\times m}$、低秩 r 以及迭代的次数 $iteration$，算法输出分解后的矩阵 $W_{n\times r}$ 和 $H_{r\times m}$，算法的核心是多次利用矩阵乘法。本文采用 MPI 消息传递模型改进矩阵乘法，进而再结合 OpenMP 共享内存模型进一步改进 NMF 算法。

4.1 基于 MPI 的并行 NMF

按照式(2)以及图 1 所示的基于分块的 NMF 迭代基本原理，基于 MPI 的并行 NMF 主要采用的是矩阵分块计算，由主进程进行控制，开启多个进程并发执行 W 和 H 的迭代计算，程序的基本流程如图 2 所示。每一轮迭代总共分为两个阶段：第一阶段是计算 H；第二阶段是计算 W。在每一轮计算完毕后，需判断是否达到循环终止条件，若达到，则计算完毕，退出程序。

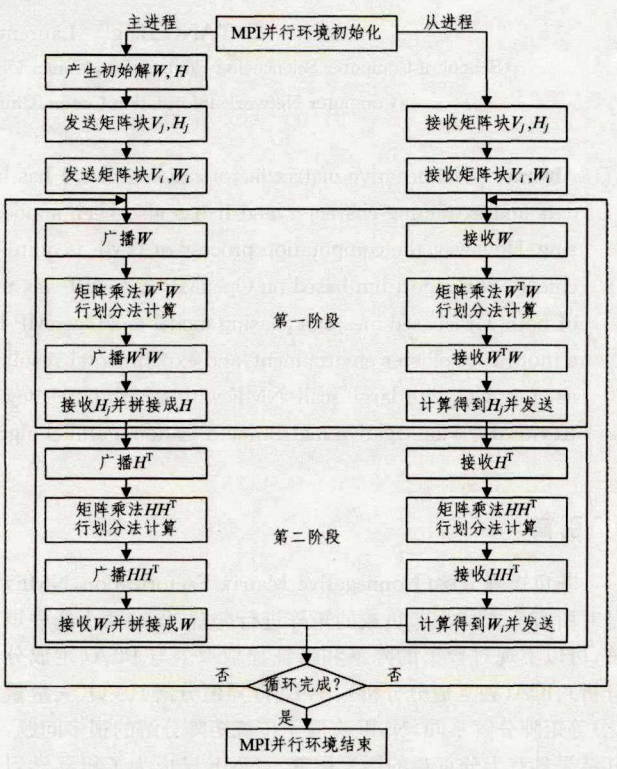


图 2 基于 MPI 的 NMF 算法的基本流程图

该算法中涉及到矩阵乘法 $W^T\times W$ 和 $H\times H^T$，可以简单地基于行划分法进行计算。以矩阵乘法 $A\times B$ 为例，行划分法的思想为：首先将矩阵 B 发送给所有从进程，然后将矩阵 A 分块，再将 A 中的数据按行分给各从进程，在从进程中计算 A 中部分行数据和 B 的乘积，最后将结果发送给主进程。

4.2 基于 MPI/OpenMP 的并行 NMF

在上述基于 MPI 的 NMF 基础之上，构建基于 MPI 和 OpenMP 的混合编程平台。在该平台上执行时能够同时实现多进程与进程内多线程编程，在执行时，每个 MPI 进程遇到一个 OpenMP 并行化的区域时便激活一组线程(自己成为主线程)。在 OpenMP 编译指导语句 # pragma omp parallel 定义的并行化区域内，进一步划分数据，并交给多个线程同时执

行。在 OpenMP 的部分，主要是实现图 2 所述的 4 个核心计算过程的多线程并行化。

1）基于行划分的矩阵乘法 $W^T \times W$；

2）迭代第一阶段 H_j 的计算：细分为一个矩阵乘法 $W^T \times V$、一个矩阵乘法 $W^T W \times H$、一个点乘和一个点除运算；

3）基于行划分的矩阵乘法 $H \times H^T$；

4）迭代第二阶段 W_i 的计算：细分为一个矩阵乘法 $V \times H^T$、一个矩阵乘法 $W \times H^T H$、一个点乘和一个点除运算。

5　算法测试环境及结果分析

5.1　测试环境

硬件环境为 8 个节点 DELL 服务器，每个节点包含 1 个 Intel Xeon E5-1650 V2 处理器（6 个计算核心，12 线程，主频 3.5GHz），32G DDR3 1886 内存，通过 1Gbps 网络进行互联。软件环境为 Ubuntu 14.04 LTS 操作系统，支持 OpenMP 和 MPI 的并行环境，采用 GCC 4.8 和 mpich2-1.4.1，程序采用 C 语言编写。在此计算平台上，为每个节点分派一个单一的 MPI 进程，每个 MPI 进程在 OpenMP 并行区域内最多发起 12 个 OpenMP 线程。

5.2　测试结果及分析

通过对串行算法（Serial-NMF）、基于 MPI 的算法（MPI-NMF）、基于 MPI 和 OpenMP 的混合算法（MPI/OpenMP-NMF）这 3 种算法的性能进行比较，来验证基于 MPI＋OpenMP 混合模型的 NMF 算法的性能及利用计算机资源的能力。

选取 4 组固定矩阵规模（阶数分别为 4800，6400，9600，12800，r 的值分别为 200，400，600，800）进行算法测试，迭代次数为 200，总共设计了 7 种测试实验（I－VII），对比了 Serial-NMF，MPI-NMF，MPI/OpenMP-NMF 这 3 种算法，节点数目有单节点、4 节点和 8 节点 3 种情况，测试实验的详细说明如表 1 所列。

表 1　测试实验的详细说明

测试实验	算法	说明
I	Serial-NMF	单节点串行
II	MPI-NMF	单节点 MPI 并行
III	MPI-NMF	4 节点 MPI 并行
IV	MPI-NMF	8 节点 MPI 并行
V	MPI/OpenMP-NMF	单节点 MPI 和 OpenMP 混合
VI	MPI/OpenMP-NMF	4 节点 MPI 和 OpenMP 混合
VII	MPI/OpenMP-NMF	8 节点 MPI 和 OpenMP 混合

首先，对运行时间进行测试，结果如表 2 所列。

表 2　运行时间/s

矩阵阶数	测试实验						
	I	II	III	IV	V	VI	VII
4800	520	309	160	89	245	78	28
6400	938	780	232	139	364	147	49
9600	1290	1020	340	195	745	206	74
12800	3250	3010	728	420	1991	510	180

然后，通过计算不同算法的加速比来衡量算法效果。加速比为单机串行方法的计算时间与并行方法的计算时间的比值。根据运行时间即可计算得到不同模型的加速比，结果如图 3 所示。

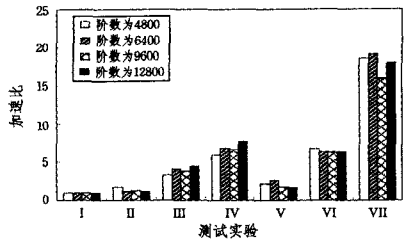


图 3　加速比计算结果

对不同阶数的矩阵进行测试，MPI/OpenMP 的加速比优势非常明显。从图 3 可以看出，MPI 和 OpenMP 混合模型的执行时间远少于 MPI 模型的并行执行时间，并且表现出非常高的执行效率。

结束语　本文实现了 *MPI＋OpenMP* 的并行非负矩阵分解算法。测试实验证明，基于 MPI 和 OpenMP 的混合模型具有较高的执行效率，更加充分发挥了节点间分布式存储和节点内共享存储的优势。在多核处理器环境下该模型可以有效提高并行计算性能，是一种高效、可行的并行编程策略，能够处理超大规模高维非负矩阵的分解。但这种层次化的并行结构也存在一些问题，例如：MPI 进程数目受限、MPI 进程间的通信带宽和延迟、OpenMP 线程产生的系统开销等问题。算法还存在有待改善的地方，例如：可改进 MPI 主进程和从进程间的传输，进一步提高效率；此外，可以对重叠计算和传输做一些改进；针对稀疏矩阵的非负分解也可以做一些优化。

参 考 文 献

[1] LEE D, SEUNG H. Learning the Parts of Objects by Nonnegative Matrix Factorization[J]. Nature, 1999, 401: 788-791.

[2] KOREN Y, BELL R, VOLINSKY C. Matrix Factorization Techniques for Recommender Systems[J]. IEEE Computer, 2009, 42(8): 40-49.

[3] CHEN Y, REGE M, DONG M, et al. Non-negative matrix factorization for semi-supervised data clustering[J]. Knowledge and Information Systems, 2008, 17(3): 355-379.

[4] CHOO J, LEE C, REDDY C, et al. UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization [J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 1992-2001.

[5] WANG W. Instantaneous Versus Convolutive Non-Negative Matrix Factorization: Models, Algorithms and Applications to Audio Pattern Separation[M] // Machine Audition: Principles, Algorithms and Systems. IGI Global, 2010: 353-370.

[6] LIAO R, ZHANG Y, GUAN J, et al. CloudNMF: A MapReduce Implementation of Nonnegative Matrix Factorization for Large-scale Biological Datasets[J]. Genomics, Proteomics & Bioinformatics, 2014, 12(1): 48-51.

[7] REN X X, TANG L, LI R F, et al. Study and Implementation of OpenMP Multi-thread Load Balance Scheduling Scheme[J]. Computer Science, 2010(11): 148-151. (in Chinese)
任小西，唐玲，李仁发，等. OpenMP 多线程负载均衡调度策略研究与实现[J]. 计算机科学，2010(11)：148-151.

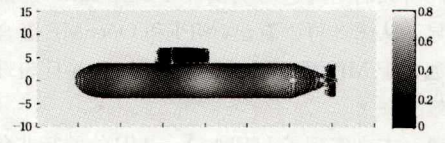
[8] FENG Y, ZHOU S Q. Research on development of mixed mode MPI+OpenMP applications[J]. Computer Systems & Applica-

tions,2006,15(2):86-89. (in Chinese)
冯云,周淑秋. MPI+OpenMP 混合并行编程模型应用研究[J]. 计算机系统应用,2006,15(2):86-89.

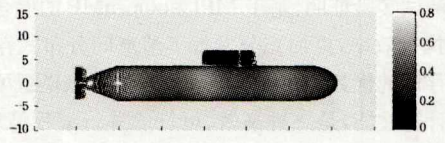
[9] PAN W,CHEN L Y,ZHANG J H,et al. Research on MPI+OpenMP hybrid programming paradigm based on SMP cluster [J]. Application Research of Computers,2009,26(12):4592-4594. (in Chinese)
潘卫,陈燎原,张锦华,等. 基于 SMP 集群的 MPI+OpenMP 混合编程模型研究[J]. 计算机应用研究,2009,26(12):4592-4594.

(上接第 50 页)

生是由于沿壳体表面传播的弹性波互相干涉导致,这也是表面声场出现亮区、暗区间隔分布的原因。

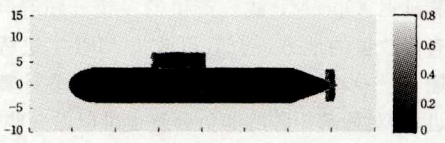


(a)非照射面

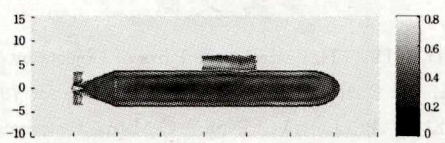


(b)照射面

图 7 入射波 230Hz 正横入射时,Benchmark 表面声场分布



(a)非照射面



(b)照射面

图 8 入射波 710Hz 正横入射时,Benchmark 表面声场分布

正横入射时,Benchmark 模型收发分置散射的目标强度(TS)如图 9 所示。

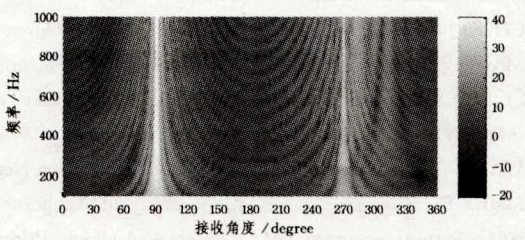


图 9 正横入射时,模型分置目标强度

正横入射时,前向散射(90°)目标强度最大,其次是反向散射(270°)方向。在前向和反向的两侧出现类似“圣诞树”的弧形散射条纹;在反向散射方向的右侧,由于目标主体部分与指挥室、尾翼的散射回波干涉,导致出现 3 条几乎与纵轴平行的散射条纹,频率越高,散射条纹越清晰。

综上,Benchmark 模型低频声散射主要是几何散射,体现为在前向和镜反射方向出现类似“圣诞树”状的散射条纹;其次,在谐振频率点处出现谐振散射,谐振散射导致表面声场出现亮区和暗区的间隔分布。

结束语 针对大尺度弹性壳体目标低频声散问题建立的有限元耦合边界元模型,本文采用并行计算技术进行求解。首先根据有限元矩阵的稀疏性与边界元矩阵的稠密性特点,分别并行计算生成系数矩阵。对于所形成的大型非对称线性方程组,采用并行化的 GMRES(m)迭代算法进行求解,并根据文献[6]中的方法对迭代后的近似值进行修正,达到加速收敛的目的。

本文以弹性球壳的声散射数值仿真为例,分析了并行计算的加速比以及迭代步数 m 对收敛精度的影响。最后计算了 Benchmark 模型的收发分置声散射特性,分析了其分置目标强度和谐振散射时的表面声场分布。

还需进一步研究的是适合求解的预处理 GMRES(m)迭代算法。

参 考 文 献

[1] EVERSTINE G C,HENDERSON F M. Coupled finite element/boundary approach for fluid-structure interaction[J]. Journal of the Acoustical Society of America,1990,87(5):1938-1947.

[2] JEANS R,MATHEWS I C. A unique coupled boundary element/finite element method for the elastoacoustic analysis of fluid-filled thin shells [J]. Journal of the Acoustical Society of America,1993,94(6):3473-3479.

[3] SAAD Y. Iterative Methods for Sparse Linear Systems(2nd Edition)[M]. SIAM,2003:171-194.

[4] NACHTIGAL N M,REICHEL L,TREFETHENK L N. A hybrid GMRES algorithm for nonsysmetric linear systems [J]. Siam Journal on Matrix Analysis and Applications,2006,13(8):796-825.

[5] AN H B,BAI Z Z. A globally convergent Newton-GMRES method [J]. Mathematica Numerica Sinica,2005,27(2):151-174.

[6] NIU Q,LU L Z,WANG R R. A Modified Gmres Method for Solving Large Nonsymmetric Linear Systems [J]. Numerical Mathematics A Journal of Chinese Universities,2005,27(S1):193-199.

[7] 朱伯芳. 有限单元法原理与应用(第三版)[M]. 北京:中国水利水电出版社,2009:192-224.

[8] BAI M R. Application of BEM(boundary element method)-based acoustic holography to radiation analysis of sound source with arbitrarily shaped geometries [J]. Journal of the Acoustical Society of America,1992,92(1):533-549.

[9] BAO X M,HE Z Y. Investigation on acoustic holography reconstruction of target scattering field [J]. Acta Acustica,2000,25(3):254-264. (in Chinese)
暴雪梅,何祚镛. 目标散射场全息重建方法研究[J]. 声学学报,2000,25(3):254-264.

[10] NELL C W,Gilroy L E. An improved BASIS model for the BeTSSi submarine[R]. DRDC Atlantic TR,2003.