

基于改进的 Map/Reduce 及模式空间划分的数据挖掘

刘 骞, 陈 明

(中国石油大学 计算机科学与技术系, 北京 102249)

摘 要: 为通过 Map/Reduce 基于键/值对以处理数据集与模式集的多对多的对应关系的方式实现数据挖掘, 解决在较复杂的类型的模式的挖掘中存在的由组合爆炸导致的模式集过大的问题, 提出了通过模式空间划分实现将处理数据集与模式集的对应关系的问题转化为处理数据集与子模式集的集合的对应关系的问题的方法, 并对 Map/Reduce 集群的调度机制及组织、处理键/值对的方式进行了改进, 增强了 Map/Reduce 执行模式挖掘任务的能力. 在 Map/Reduce 集群上采用该思路实现某些较复杂类型的模式的挖掘算法时的并行度高于 Map/Reduce 化的传统算法.

关键词: 改进的 Map/Reduce; 集群; 多对多映射; 模式空间划分; 模式编码; 数据挖掘

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2011)08-0140-03

Data Mining Based on Improved Map/Reduce and Pattern Space Division

LIU Qian, CHEN Ming

(Department of Computer Science and Technology, China University of Petroleum, Beijing 102249, China)

Abstract: In order to realize data mining through map/reduce based on key/value pairs, the way of processing the many-to-many corresponding relationship between the data set and the pattern set is adopted. For some of the more complex types of pattern, the pattern set is so large because of combinatorial explosion that the corresponding relationship cannot always be processed directly by the map/reduce. Therefore a way of pattern space is proposed to convert the problem of processing the many-to-many corresponding relationship between the data set and the pattern set to the problem of processing the many-to-many corresponding relationship between the data set and the set of the sub-pattern sets. At the same time, the scheduling mechanisms of the map/reduce cluster and the way of organizing the key/value pairs is improved to enhance the ability of map/reduce to execute pattern mining tasks. The results show that higher parallelism is achieved on map/reduce clusters by using this idea than the map/reduce of the traditional algorithm in mining some of the more complex types of pattern.

Key words: improved map/reduce; cluster; the many-to-many corresponding relationship; pattern space division; pattern coding; Data Mining

1 引言

当前对通过大规模 Map/Reduce 集群进行海量数据分析的研究大都遵循将传统数据挖掘算法进行 Map/Reduce 化^[1-2], 从而将其迁移到云平台上的思路. 由于受传统算法建模方式的限制, 该方法有时不适用于在云集群上实现复杂类型模式的挖掘. 而信

息检索等领域中通过 Map/Reduce 基于键/值对数据集进行的处理, 本质上是对初始键与结果键之间的多对多的对应关系的处理. 本文在基于 Google 阵营的数据分析型集群技术的 HADOOP 开源框架基础上, 提出通过 Map/Reduce 基于键/值对以处理数据集与模式集的多对多的对应关系的方式进行数据挖掘的思路.

2 Map/Reduce 与模式空间划分

Google 的云计算技术源于信息检索领域的大规模数据分析任务^[3],这类任务的关键是通过 Map/Reduce 以键/值对的方式分析处理数据集中的记录及整个模式空间中的模式间的多对多的对应关系。

但是对于复杂类型的模式如频繁模式、序列模式等的挖掘,由于这类问题本身的非线性结构决定了其模式空间中可能的目标模式的海量性,直接通过 Map/Reduce 的映射、规约机制分析数据集与海量模式集的对应关系来得出目标模式是不现实的。

故采取将非线性问题线性化的思路对复杂模式挖掘问题进行转化,使得每个从节点需要处理的模式集不至过大并且其数据分片所蕴含的模式与其他节点是不相交的。即从问题本身的结构出发,根据具体数据集的特点构造模式空间,并选择最合理的模式空间划分策略的方式来实现在整个模式空间的各分划中分别无损的挖掘出目标模式集的子集,最后合并所有以上子集从而得到整个目标模式。

3 改进 Map/Reduce 并行编程框架

3.1 代码本地化

进行复杂类型模式的挖掘时,需要在用户程序控制下运行多轮的 Map/Reduce 任务。而 Map/Reduce 并行框架采用函数式编程方式,每轮 Map/Reduce 调度要首先将用户定义的 Map 函数和 Reduce 函数编译生成的文件分发到大量的集群节点上,这个过程会导致整个任务效率的降低。为提高调度效率,在 Map/Reduce 中加入组播控制机制,并将挖掘目标模式的各轮 Map/Reduce 任务的程序文件加入到各集群节点的 HADOOP 可执行程序路径中,并在各集群节点上建立消息监听器,当接收到主节点的组播控制消息时可以快速启动相应 Map 或 Reduce 任务。该机制不仅大大提高了处理子系统的调度效率而且降低了网络子系统的负载。

3.2 组播全局数据特征信息

根据整个数据集的特征来指导、优化各个任务对本地数据集的处理是十分必要的。这就要求用户程序能够将控制集群提取出全局特征信息,并进行处理、组织后通过主节点加以并分发到参与任务的从节点作为从节点的全局变量指导从节点上的所有任务。该机制以较小的代价弥补了 share-nothing 体系结构^[4]的不足,扩充了 Map/Reduce 对复杂任务

的处理能力。

3.3 结构化的键/值对组织

为了实现模式空间的构造、划分,需要在全局特征数据的指导下将子数据集中的蕴含的模式分别映射到各个子模式空间中去。而不同节点上的数据子集向子模式空间映射的过程就是建立中间结果键/值对的过程,为了降低规约过程中的网络负载及提高合并各 Map 任务对应的中间结果键/值对的效率,这里的值的数据类型应采用高效的数据结构来实现。该机制是 Map/Reduce 能够高效处理复杂类型模式的关键,不同的模式空间构造、划分策略决定了具体数据结构的选择。图 1 描述了该机制下以结构化键值对的方式对数据进行组织处理的原理。

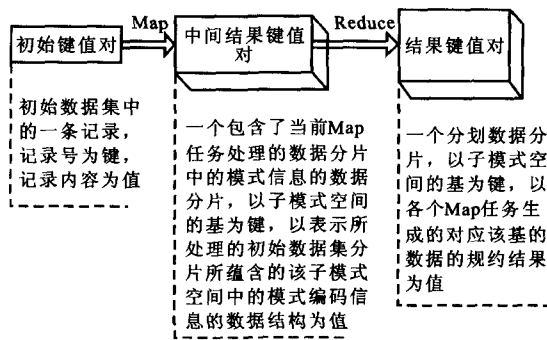


图1 键值对的组织、处理原理图

4 模式空间构建及划分策略的制订

4.1 构建模式空间

对于较复杂的类型的模式,已有传统算法难以直接基于 Map/Reduce 实现并行处理。为了充分利用 Map/Reduce 集群强大的并行处理能力,提高这些类型的模式的挖掘效率,需要根据 Map/Reduce 的处理特点对这些类型的模式的挖掘问题本身重新建模,构建模式空间并制定合理的模式空间划分策略,便于在划分后的各个子模式空间中并行地进行子模式的挖掘。

模式空间的构建遵循在尽可能控制整个任务总的处理开销、存储开销、网络负载不至过大的前提下能够提供可变尺度、可变粒度的灵活划分策略,从而可以根据具体任务的工作量及集群中节点的处理能力自动实现划分策略的选择及执行。

4.2 模式编码

为了方便描述及实现模式空间的构建、划分,引入模式编码的概念。从不同的角度出发一种类型的模式的可以有多种不同的编码方案,对应多个不同模式空间。在定义了模式编码的基础上,根据具体模

式的特点进一步在编码的基础上定义诸如模式长度、距离、模式空间的基等概念, 基于这些概念来制订多样化的模式空间划分策略, 从而将原始数据集映射到满足不同编码特征的子模式空间中去, 根据各个子模式空间的编码特征分别进行挖掘。

4.3 划分模式空间

初始数据集中的每个分片都部分地蕴含了结果模式集中的模式, 所以无法分别在每个分片中直接进行目标模式的挖掘。通过对初始数据集进行处理, 将其中蕴含的模式信息分别映射到互相独立的子模式空间中去以得到分别完整地蕴含各个子模式空间中的模式的分划数据集是划分模式空间的目的所在。分划数据集的每个分片对应一个子模式空间, 位于集群的一个节点上。图 2 示意了初始数据集有 2 个分片, 划分为 3 个子模式空间及生成相应的 3 个分划数据分片的情况下该过程中的模式信息的转移汇聚情况。

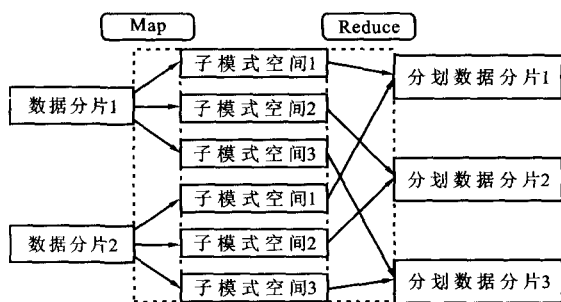


图 2 模式信息转移汇聚过程图

划分的模式空间的个数应根据总的处理任务、集群中节点的个数、每个节点的处理能力等灵活地确定。而实现该点的关键就是提取出全局数据集的特征信息, 在此信息指导下确定模式空间的基的个数亦即子模式空间的粒度。基的选取决定了子模式空间的粒度及对应的各个分划数据集上进行子模式挖掘的时间、空间复杂度。

在通过一轮 Map/Reduce 提取模式空间的特征信息后, 用户程序通过分析算法对其进行分析, 确定模式空间的编码方案、模式空间的基及编码后的数

据集到子模式空间的映射方式。用户可以通过主节点将这些分析出的信息及特征信息本身组播到所有从节点, 实现 share-nothing 体系结构下全局信息的高效访问。在这些全局信息的指导下, 各从节点并行地对数据分片进行编码、映射等操作。

5 子模式空间上的模式挖掘

模式空间划分完成后, 子模式空间中的模式被分划数据集的某个数据分片唯一地蕴含, 设计合理的算法在全局数据的特征信息的指导下根据各个子模式空间的特征分别在相应的数据分片上并行地进行模式的挖掘。

6 结束语

本文在深入分析 Map/Reduce 并行编程框架进行数据分析的特点的基础上, 提出在对复杂类型模式挖掘的问题本身重新建模的基础上, 通过实现模式空间划分来实现在各个子模式空间上并行地进行模式的挖掘的方法。与通过 Map/Reduce 化将传统算法迁移到云集群上相比, 本文提出方案具有更广泛的适用性并可取得更高的并行化程度。

参考文献:

- [1] Ng Andrew Y, Bradski Gary, Chu ChengTao. MapReduce for machine learning on multicore [R]. NIPS, 2006.
- [2] Dan Gillick, Arlo Faria, John DeNero. Map/Reduce: distributed computing for machine learning[R]. 2006.
- [3] Dean Jeffrey, Ghemawat Sanjay. Map/Reduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1):107-113.
- [4] 陈康, 郑伟民. 云计算: 系统实例与研究现状[J]. 软件学报, 2009, 20(5):1337-1348.

作者简介:

刘 赛 男, (1985-), 博士研究生, 研究方向为分布式计算、云计算、软件工程。

(上接第 139 页)

- [2] 张赛男, 殷兆麟. 软件系统 UML 建模与其安全建模的集成[J]. 计算机工程, 2007, 33(8):86-88.
- [3] Pende T. UML 宝典[M]. 北京: 电子工业出版社, 2004:146-152.
- [4] Brose G, Koch M, Lohr K P. Integrating access control design into the software development process [C]//

Proc. of the 6th International Conference on Integrated Design and Process Technology. Trondheim, 2002.

作者简介:

张赛男 女, (1979-), 硕士, 讲师. 研究方向为软件工程、网络安全。