文章编号:1671-4598(2014)05-1510-03　　中图分类号:TP311　　文献标识码:A

基于云计算的并行 K-means 聚类算法研究

谢雪莲，李兰友
（南京工程学院 信息化建设与管理办公室，南京　211167）

摘要：目前数据呈爆炸式增长，海量存储状态，给聚类研究带来了诸如计算复杂性和计算能力不足都很多问题；而云计算平台通过负载均衡，动态配置大量的虚拟计算资源，有效地突破了耗时耗能的瓶颈，在海量数据挖掘中体现出了其独特的优势；文章深入研究了基于云计算平台 Hadoop 的并行 K-means 算法，并结合 MapReduce 分布式计算模型，给出了算法设计的方法和策略，包括 MapReduce 处理的 map、shuffle 和 Reduce 3 个过程，仿真结果表明 K-means 并行算法的效率较高。

关键词：云计算技术；Hadoop；MapReduce；K-means 算法

Research on Parallel K-means Algorithm Based on Cloud Computing Platform

Xie Xuelian，Li Lanyou
（Information Construction and Management Office，Nanjing Institute of Technology，Nanjing　211167，China）

Abstract：Now with the explosive growth of data，there are a lot of problems such as complexity and lack of computing resources in cluster research . The cloud computing platform breaks the state of high energy—consuming and shows the advantages in the massive data mining through load balance and dynamic configuration of virtual computing resources. The parallel K-means algorithm based on Hadoop platform was designed . The algorithm methods and strategies were shown combined with MapReduce distributed computing model including map，shuffle and Reduce. Simulation results show that the parallel K-means algorithm has strong computing efficiency.

Key words：cloud computing；Hadoop；MapReduce；K-means algorithm

0　引言

数据这个词语，现代社会已经不再陌生，它扎根于我们的信息社会的最底层。数据的增长不再是缓慢的、线性的，它呈现的是爆炸式的、海量型的。各个行业中产生的数据洪流，十分繁杂但又极具价值，要从中发现有用的信息和知识，为未来的决策提供支持，就是聚类算法研究的目标。

“物以类聚，人以群分”，聚类算法[1]就是研究如何将相似元素集合成类或簇的过程。不过，传统聚类算法在面对海量数据时，在时间复杂性和空间复杂性上遇到了问题，而基于云计算的并行聚类算法是解决此问题的有效途径。

云计算[2-3]是一种新的计算模式，如亚马逊、百度、腾讯等大公司等竞相推出自己的云计算产品。如腾讯推出的“微云”云存储服务，从 2013 年 9 月 2 日起，每个 QQ 号可按照活动规则免费获得 10T 容量，并且上传资源时，利用“极速秒传”技术快速查询所上传资源是否已经在云中，如果有，则不再进行上传，而是以“极速妙传”的形式显示上传成功。不管是 10T 的超大容量网盘，还是近 3G 的视频资源在 5 s 内秒传成功，这在以前都是无法想象的。Hadoop[4]是一个开源的易开发和并行处理的云计算平台，它包括两部分：Hadoop 分布式文件系统（HDFS）和 MapReduce 计算模型。利用 Hadoop 平台处理海量数据的优势，构建一种改进的并行 K-means 聚类算法，在 Hadoop 平台上实现 K-means 算法的 MapReduce 并行化，实验表明 K-means 并行算法的效率较高。

1　云平台技术

云计算是一种通过 Internet 以服务的方式提供动态可伸缩的虚拟化的资源计算模式。Hadoop 是云计算应用的一个开源平台，用户可以充分利用 Hadoop 云计算平台的海量计算和存储能力，在不了解分布式底层细节的情况下，完成程序的分布式执行。Hadoop 具有很多重要的特性，首先，可扩展能力强，集群数量可以方便地扩展到数以千计的节点，能可靠地存储和处理 PB 级的数据；成本低，可以通过普通机器组成集群来分发和处理数据；高效率，能够在节点之间动态的分发数据，保证各个节点的动态平衡，并在数据所在的节点上并行的处理，使得处理的速度非常快；可靠性高，能自动地维护数据的多份复制，并且在任务失败后，能够自动地重新部署计算任务。Hadoop 在海量数据存储、筛选、加工和挖掘方面表现出良好的性能，并能提供稳定可靠的接口，可以构建一个具有高可靠性和良好扩展性的分布式系统[5]。Hadoop 中包含很多子项目，如图 1 所示。

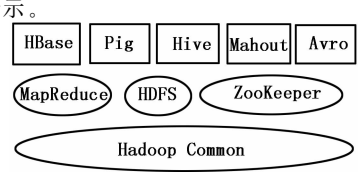

图 1　Hadoop 子项目结构图

收稿日期:2013-12-20；　**修回日期**:2014-02-20。

基金项目:南京工程学院校级科研重点基金(QKJA2011011)；江苏省属高校自然科学基金(11KJB510005)；国家自然科学基金(61104085)。

作者简介:谢雪莲(1977-)，女，江苏无锡人，工程师，硕士，主要从事数据库安全方向的研究。

Hadoop Common：它是整个 Hadoop 项目的核心，Hadoop 体系最底层的一个模块，定义了一系列组件和接口，并且提供所需 API。

MapReduce：是一种高效的分布式编程模型，它将数据处理分为 Map 和 Reduce 两个阶段。MapReduce 将 key/value 键值对作为输入和输出形式，用户指定 key 和 value 的类型，同时用户自定义 Map 和 Reduce 函数。Map 过程中将数据以 key/value 键值对形式作为输入和输出，Map 过程产生 key/value 键值对的中间结果，将同一 key 的 value 值进行合并，Reduce 过程将合并后的 key/value 键值对作为输入进行处理，最后输出用户需要的 key/value 键值对形式。MapReduce 的运行流程如图 2 所示。

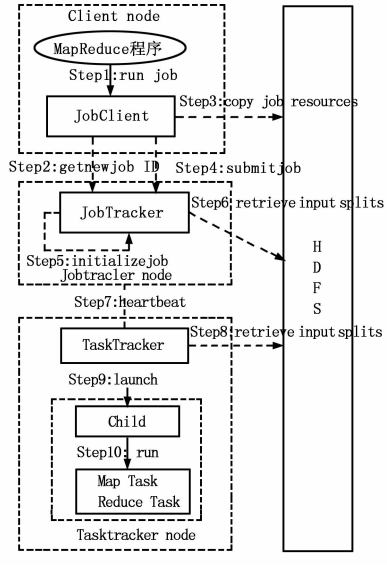

图 2　MapReduce 运行流程图

整个模型的最上层有 4 个实体，客户端（client）主要负责向 MapReduce 框架提交作业；作业追踪器（JobTracker）全权负责调度作业的运行；任务追踪器（TaskTracker）负责运行输入切片数据，负责具体任务的执行；分布式文件系统（HDFS）提供实际的存储服务，用于向所有的节点共享作业所需的资源[6]。流程步骤如下：

（1）通过 JobTracker. getNewJobId（）方法向 JobTracker 请求当前作业的 ID（步骤 1）；

（2）获取各个路径信息时会检查作业的对应路径；

（3）计算作业的输出划分，并将划分信息写入分块文件（步骤 2）；

（4）将运行作业所需的资源，包括计算所得的输入划分和作业的配置文件，复制到作业对应的 HDFS 中（步骤 3）；

（5）通过调用 submitJob（）方法来提交作业，并告知 JobTracker 准备执行作业（步骤 4）；

（6）作业调用 JobTracker. submitJob（）方法后，JobTracker 会将此调用放入内部的变量中，然后采用先进先出方式进行调度（步骤 5）；

（7）JobTracker 首先从 HDFS 中获取作业对应的分块信息（步骤 6），为分配 Map 任务做准备，然后创建并初始化 Map 和 Reduce 任务，同时指定任务 ID；

（8）TaskTracker 每隔一段时间向 JobTracker 发送心跳，告诉 JobTracker 此 TaskTracker 是否存活，是否准备执行新任务；

（9）TaskTracker 通过 localizeJob（）方法完成任务本地化（步骤 8），然后，launchTask（）方法会为任务创建本地工作目录，并把 JAR 文件中的内容解压到这个文件夹下（步骤 9），最后启动 TaskRunner 来执行任务；

（10）最后，TaskRunner 又会启动一个新的 JVM 来运行每个任务（步骤 10）。

2　基于 MapReduce 的并行 K-means 聚类算法设计

MacQueen 提出的 K-means[7-8] 算法是一种基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度越高。

假定本数据集 $X=\{x^1, x^2, \cdots, x^N\}$，其中 x^i（$i=1, 2, \cdots, N$）表示第 i 个样本，N 为样本总数. 以 $\{m^k\}_{k=1}^K$ 表示第 k 个簇的质心，C^k 表示第 k 个簇的样本集，则 K-means 的目标函数可表示为：

$$J_{\text{k-means}} = \min \sum_{k=1}^{K} \sum_{x^i \in C^k} | x^i - m^k |^2$$

则 K-means 算法的串行执行过程为：

（1）从数据集中选择 k 个数据对象作为初始的聚类中心；

（2）根据各个聚类中心，把除聚类中心的其余所有数据对象分配到相似度最高的聚类；

（3）把所有的数据对象分配到相应的聚类后，根据每个聚类中所有数据对象的矢量重新计算每个聚簇的聚类中心；

（4）计算误差平方和准则函数，对比较连续两次的计算值，如果在设定的差值范围内，则聚类准则函数收敛，否则循环执行第（2）步和第（3）步。

K-means 算法[9-10] 的 MapReduce 并行化的思路就是把串行 K-means 算法的每次迭代转化为一次 MapReduce 计算，且可以独立操作的计算并行实现，包括样本与聚类中心的距离计算和新的聚类中心的局部计算。算法实现的过程主要包括 Map 过程，Combine 过程和 Reduce 过程，其算法流程图如下图 3 所示。

2.1　Map 函数设计

Map 函数将文件的每行作为一个样本，以 key/value 键值对的形式表示，并计算每个样本到各个聚类中心的距离，选择距离最小的聚类中心，并把该数据样本分配到该聚类中，并把该数据样本标记为所属的新聚类类别，形成 key/value 键值对的输出形式。输入数据以＜行号，记录行＞的形式表示＜key，value＞对，map 函数输入的 key 是当前记录相对于输入数据文件起始点的偏移量，value 是当前记录的各维坐标值；输出中间结果以＜聚类 id，记录属性向量＞的形式表示＜key′，value′＞对；输出数据同样为＜key′，value′＞对，key′表示聚类 id，value′表示与该聚类中心最为相似的数据对象。

Map 函数的伪代码如下：

```
map(<key,value>,<key',value'>)
{定义 input 数组,记录从 value 解析出来的每个样本的各维值 tmpLength=Math.max(); 辅助变量 tmpLength 初始化为最大值;
index=-1;index 初始化为-1;
for(int i=0;i<=k-1;i++) do { dis=input 与第 i 个聚类中心各维的距离; if(dis<tmpDis){tmpLength=dis; index=i;} }
将 index 作为 key' ;将各维坐标值作为 value' ;
输出<key',value'>; }
```

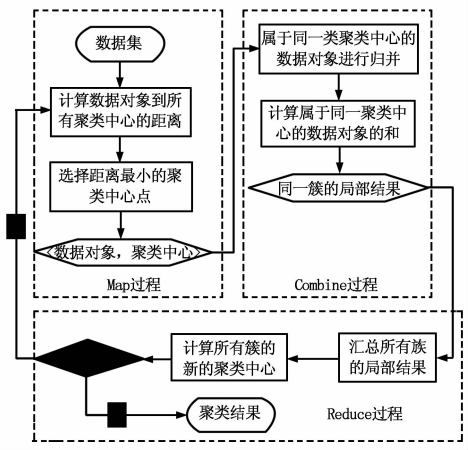

图 3 K-means 算法 MapReduce 流程图

2.2 Combine 函数设计

Combine 函数的作用就是对 Map 过程产生大量中间结果进行本地化 Reduce 处理，可以减轻数据在节点之间的传输时间耗费和带宽占用。Combine 函数的任务就是将所处节点内的 Map 结果进行预处理，对具有同一 key 值的 value 值进行处理，然后将处理得到的局部聚类结果，再传给集群中 Reduce 函数进行规约操作。Combine 函数的伪代码如下：

```
combine(<key,value>,<key′,value′>)
{初始化变量 num=0,用于统计相同聚类的记录数目
while(value.hasNext())
{ 解析 value 中每个记录的各维坐标值；
将各维坐标值累加存放到数组中；num++;}
将 key 赋值为 key′
构造存储 num 和数组信息的字符串,作为 value′
输出<key′,value′> ;}
```

2.3 Reduce 函数设计

Reduce 函数通过汇总 Combine 函数得到的局部聚类结果计算出新的聚类中心，并将其用于下一轮迭代运算。Reduce 函数首先计算每个节点输出的局部聚类结果的样本个数，并解析每个样本的各维坐标值，然后将对应的各维累加值分别对应相加，再除以刚才计算得到的总样本个数，计算结果就是新的聚类中心坐标。Reduce 函数的伪代码如下：

```
reduce(<key,value>,<key′,value′>)
{初始化 num1=0,用于统计相同聚类的总的记录个数；
while(value.hasNext())
{解析 value 中各维坐标值和记录个数 num；
各维坐标值累加;num++;}
新聚类中心=(数组中的各个分量/ num1)；
将 key 作为 key′ ;构造新各维坐标值的字符串,作为 value′ ;
输出<key′,value′> ;}
```

执行完 Reduce 任务之后，将输出结果计算新的聚类中心，并更新到 HDFS 分布式文件系统中，并将该文件复制到集群中的所有节点上，然后循环迭代计算 MapReduce Job 的误差平方和准则函数，若差值小于设定的值，则聚类准则函数已收敛，算法结束；否则，将新的聚类中心替换原来的聚类中心，启动新一轮迭代计算，同样是 Map 任务，Combine 任务和 Reduce 任务的流程，当迭代输出趋于稳定收敛时，就可以得到最终的聚类结果。

3 实验仿真

利用基于 MapReduce 的并行 K-means 聚类算法对每月的流量信息进行聚类分析，首先将文件分割的 splits 映射成一个一个<key，value>键值对，key 存储的是文件中的行偏移量，value 存储的是文本文件中的一行[11]。Map 阶段结束后，进入 Combine 阶段。Combine 过程主要是对 Map 输出<key，value>键值对，按照 key 的 hash 值进行排序和分组，最后输出到中间结。Reduce 阶段接收来自于 Combine 阶段的中间结果，并遍历整个列表并找出最大值，这个结果就是每月中的最高流量。图 4 给出了 MapReduce 的处理执行过程，包括 map、shuffle 和 Reduce 3 个过程，中间的 shuffle 过程的主要作用是将具有相同的 key 值的中间结果交给同一个 Reduce 函数去处理。

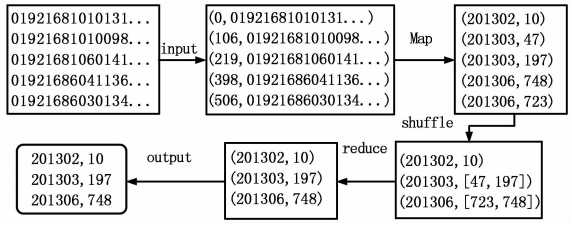

图 4 实验仿真数据处理图

4 结束语

本文介绍了云平台 Hadoop 的 MapReduce 分布式编程模型，并描述了 K-means 聚类算法的思想与流程，并给出了基于 MapReduce 的并行 K-means 算法的思路与实现过程，列出了 Map 函数、Reduce 函数和 Combine 函数的伪代码设计，并给出了实验仿真结果。不过，实验中用的数据较为单一，针对复杂的网络舆情 web 数据挖掘方面，如何应用基于 MapReduce 的 K-means 算法进行有效挖掘，还是未来需要研究的重点。

参考文献：

[1] 孙吉贵，刘 杰，等. 聚类算法研究 [J]. 软件学报，2008，19 (1)：48－61.
[2] 方 巍，文学志，等. 云计算：概念、技术及应用研究综述 [J]. 南京信息工程大学学报（自然科学版），2012，4 (4)：351－361.
[3] 陈 康，郑纬民. 云计算：系统实例与研究现状 [J]. 软件学报，2009，20 (5)：1337－1348.
[4] 赵卫中，马慧芳. 等. 基于云计算平台 Hadoop 的并行 K-means 聚类算法设计研究 [J]. 计算机科学，2011，38 (10)：166－168.
[5] 张石磊，武 装. 一种基于 Hadoop 云计算平台的聚类算法优化的研究 [J]. 计算机科学，2012，39 (10)：115－118.
[6] 穆瑞峰，苗国义. 基于粒子群优化的模糊 K-means 目标分类算法 [J]. 计算机测量与控制，2013，21 (5)：1266－1268.
[7] 郝春梅，吴 波. 基于粒子群改进的 K-means 算法 [J]. 计算机测量与控制，2013，21 (4)：1011－1013.
[8] 江小平，李成华. K-means 聚类算法的 MapReduce 并行化实现 [J]. 华中科技大学学报（自然科学版），2011，39 (supI)：120－124.
[9] 应 毅，任 凯，曹 阳. 基于改进的 MapReduce 模型的 Web 挖掘 [J]. 科学技术与工程，2013，13 (5)：1205－1209.
[10] 於跃成，王建东，郑关胜，等. 基于约束信息的并行 K-means 算法 [J]. 东南大学学报（自然科学版），2011，41 (3)：505－508.
[11] 胡爱娜，蔡晓艳. 基于 MapReduce 的分布式期望最大化算法 [J]. 科学技术与工程，2013，13 (16)：4603－4606.