# 改进的 Apriori 算法的研究及应用

刘丽娟

（南京师范大学泰州学院 信息工程学院，江苏 泰州 225300）

**摘 要**：针对传统的关联规则数据挖掘算法 Apriori 在处理海量数据时效率低、扩展性差等问题，提出利用 Hadoop 平台的编程模型 MapReduce 实现 Apriori 算法并行化的方法，通过改变数据集的大小、最小支持度、最小置信度、节点数目等几个实验对算法性能进行测试。测试结果表明，经过改进后的 Apriori 算法，具有良好的并行扩展能力，能够满足大数据处理的需求。实例分析将该算法应用到学生成绩数据中，验证了其有效性，能更好地为教育决策服务。

**关键词**：云计算；关联规则；数据挖掘；大数据；学生成绩

**中图法分类号**：TP311 **文献标识号**：A **文章编号**：1000-7024（2017）12-3324-05

**doi**：10.16208/j.issn1000-7024.2017.12.023

# Research and application of improved Apriori algorithm

LIU Li-juan

(College of Information Engineering, Nanjing Normal University Taizhou College, Taizhou 225300, China)

**Abstract**: To solve problems such as low efficiency, poor scalability of the traditional association rule data mining algorithm Apriori in dealing with massive data, programming model MapReduce on Hadoop platform was used to realize the parallelization of Apriori algorithm. By changing the size of several experimental data sets, minimum support, minimum confidence, the number of nodes and other experiments were used to test the performance of the algorithm. The results show that the improved Apriori algorithm has good parallel spreading ability and it can meet the requirements in huge data processing. The improved algorithm was applied to the case study to process the student achievement data, verifying the validity of the proposed algorithm, which better serves the educational decision.

**Key words**: cloud computing; association rules; data mining; large data; student text score

## 0 引 言

大数据经历了批处理、实时处理、批/实时混合处理三代技术更新[1]，截止目前，Hadoop 平台仍然被人们用来和新型技术一起处理数据。纵观国内外大量文献，尽管大数据在各行各业应用广泛，但其在教育中应用很少，主要在教学模式、管理稍有涉及，部分研究者利用传统的 Apriori 进行教育数据的挖掘，但其弊端是当数据量增大到一定程度时算法效率较低。后来也有一些研究者将 Apriori 和大数据技术结合进行了改进，但很多都是对 Apriori 生成频繁项集作了修改，并没有实现强关联规则。在此基础上，本文提出借助大数据技术的优势来改进 Apriori，利用目前流行的云计算分布式平台 Hadoop 为大数据提供很强的存储能力和计算处理能力，针对传统 Apriori 算法的局限性，搭建 Hadoop 计算机集群，利用 MapReduce 对算法进行改进，最后将并行化的算法用来对高校学生的成绩进行分析、处理，找出课程之间的隐含规则，提高传统 Apriori 算法在处理大数据时的效率，更好地为教育服务。

## 1 Apriori 算法简介

Apriori[2]是对现在影响较大的数据挖掘算法，利用迭代形式，依次循环处理，直到生成最终的频繁 K 项集，整个过程需要多次扫描数据库，最后根据频繁 K 项集分析总结出强关联规则，供用户决策分析。

Apriori 算法存在以下几个主要问题：多次迭代时，每次都需要遍历数据库，非常繁琐；计算候选项集[3]时会占用很大的内存容量，进而会导致内存不够，整个程序停止运行；寻找频繁项集需要消耗大量时间。这些缺点导致

收稿日期：2017-03-06；修订日期：2017-05-16

基金项目：江苏省高校自然科学研究面上基金项目（16KJB140007）；泰州市科技支撑（软科学）计划基金项目（RM201429）

作者简介：刘丽娟（1983-），女，江苏泰州人，硕士，讲师，研究方向为云计算、大数据。

E-mail：149110375@qq.com

Apriori 算法的执行效率极低。

## 2　基于 Hadoop 的 Apriori 算法的改进

### 2.1　Hadoop 平台[4]简介

云计算是依赖于互联网、能向用户提供各种软硬件平台和信息的一种进行基础设施交付、使用的新技术，Hadoop 技术是现如今广泛得到关注、以 HDFS(Hadoop distributed filesystem)[5]和 MapReduce[6]为核心的云计算平台。

分布式文件系统 HDFS 非常适合应用在大规模数据集中。MapReduce 是对大规模数据进行并行计算的一种编程模型和计算框架，封装了并行处理、数据分布、负载均衡等复杂的细节，使并行编程模型变得更简单。使用 MapReduce 框架编写程序，只要实现 Map 函数和 Reduce 函数[7]即可。

### 2.2　Apriori 算法的 MapReduce 改进

在处理海量数据时，Apriori 算法在执行过程中效率较低，利用 MapReduce 对其进行更改，主要是利用 Map() 和 Reduce() 来从海量数据中快速寻找出频繁项集，使 Apriori 算法处理海量数据不再受到单机运算能力的限制。

改进 Apriori 算法的基本思路分为以下几步：

(1) 在 Map 阶段主要利用 k-1 项集的连接操作得出 k 项集，以项集记为 key，支持度记为 value；

(2) 在 Reduce 阶段主要将相同 key 的 value 合并，并且进行剪枝操作，筛选符合要求的项集。

(3) 由频繁项集生成关联规则。

改进 Apriori 算法的流程如图 1 所示。

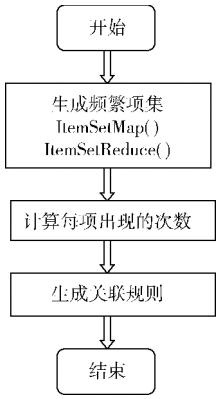


图 1　改进 Apriori 算法流程

改进 Apriori 算法的伪代码如下所示：

(1) 生成频繁项集

```
ItemSetMap( )
输入：数据集 D
输出：项目集
for each m in D do
  for each item in m do
    context. write(item，1)
  end
end
ItemSetReduce( )
输入：<item，1>，最小支持度 minsup
输出：频繁项集 freqItemSets
int s=0；
for each v in values do
  s=s+value. get( )
  if(s>=minsup)
    context. write(key，s)
  end
end
```

(2) 生成关联规则

```
输入：频繁项集 freqItemSets，最小置信度 minconf
输出：k 项频繁集存入路径 k __outputpath
for each itemset in freqItemSets
  for all i-itemset
    conf=1. 0*itemset. getSupport( ) /
      occurrence. get(itemset. getItems())
    if(conf>=minconf)
      context. write(itemset，conf)
    end
  end
end
```

## 3　改进算法的分析

### 3.1　Hadoop 分布式环境的搭建[8]

当利用传统的 Apriori 算法进行数据挖掘时，数据量增大到一定程度，计算机内存将无法支撑算法正常执行。为了运行大数据集，需要搭建 Hadoop 计算机集群。

#### 3.1.1　配置 Hadoop 实验平台

本实验在 VMware workstation 上设置 5 台虚拟机作为 Hadoop 集群环境，1 台作为 NameNode，另 4 台作为 DataNode。

#### 3.1.2　Hadoop 实验平台的搭建

(1) 安装虚拟机，配置 Linux 系统。

(2) 配置 ssh(secure shell) 实现 Hadoop 节点间用户的无密码访问。

(3) JDK(Java development kit) 的安装配置，Hadoop 采用 java 编写，需要安装 JDK。

(4) 下载 Hadoop 软件安装，并进行配置，启动成功。

#### 3.1.3　安装 eclipse

下载 eclipse 软件到 Hadoop 目录，并解压缩到此目录，进行文件的复制，重启 eclipse 即可。

至此 Hadoop 的平台基本搭建完成。

### 3.2 实验结果分析

（1）并行挖掘结果验证

实验数据使用我校学生成绩的历史数据，经过改进后的并行算法对应程序与原始的串行算法对应程序两者的挖掘结果来进行对比，如果结果一致，表明改进后的算法是可靠的，否则表明其设计有误，无法进行正确的数据挖掘。表 1 显示了不同大小的文件在两种算法下最终的挖掘结果。

表 1 文件挖掘结果

| 文件大小 | 算法 | FIM1 | FIM2 | FIM3 | FIM4 |
|---|---|---|---|---|---|
| 100 M | 串行 | 20 | 43 | 37 | 10 |
| 100 M | 并行 | 20 | 43 | 37 | 10 |
| 250 M | 串行 | 24 | 45 | 39 | 12 |
| 250 M | 并行 | 24 | 45 | 39 | 12 |

表 1 中结果显示，在不同挖掘数据文件大小情况下，分别利用并行算法和串行算法挖掘出来的结果是一致的，从频繁项集 FIM1 到 FIM4 的结果也相同，因此，本文使用的改进后的算法是可靠的，能够准确挖掘出频繁项集。

（2）并行与串行算法性能对比

分别用并行程序和串行程序对大小为 50 M、100 M、150 M、200 M、250 M、300 M、350 M、400 M 的数据进行挖掘，实验结果如图 2 所示。

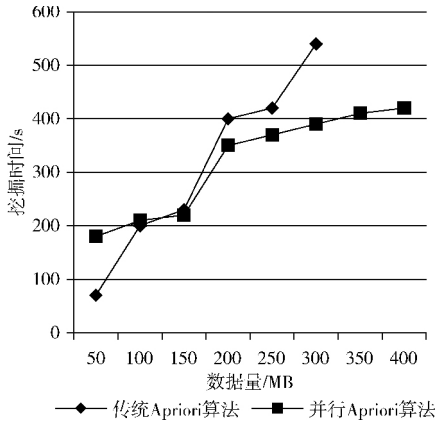


图 2 并行与串行算法性能对比

图 2 中可以得出结论，当数据量较小时，由于 Hadoop 集群调度分配任务等需要消耗一定时间，并行算法执行时间反而更长，但随着数据量的增大，其优势体现出来，执行时间较短。串行方式下单机使用传统的 Apriori 算法当数据规模增大到一定量时，最终导致机器内存溢出，程序运行失败。因此，在处理海量数据时，改进的并行算法能够利用集群优势，将数据分片处理，大大降低了单一节点的内存消耗，较好地完成任务。

（3）改变支持度[9]、置信度时两种算法性能分析

改进后的算法其可靠性已经在实验（1）中得到了验证，下面分别通过改变两种算法的最小支持度、最小置信度来验证并行算法的稳定性，图 3 为改变支持度时的响应时间对比，图 4 为改变置信度时的响应时间对比。

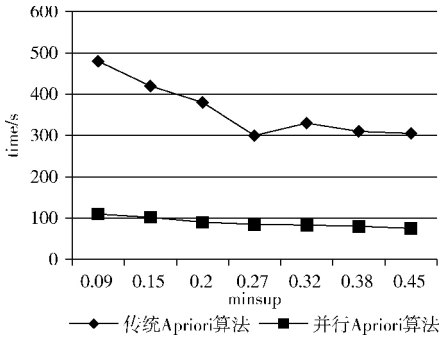


图 3 改变支持度的响应时间对比

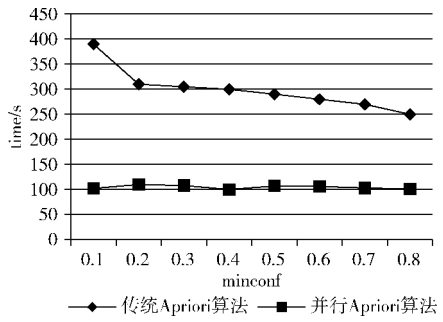


图 4 改变置信度的响应时间对比

从图 3、图 4 的实验分析结果可以得出结论，改进后的并行算法在数据挖掘时不仅可行，并且性能十分稳定。

（4）节点数目变化对挖掘的影响

在搭建的 5 台虚拟机 Hadoop 集群中，其中一台机器作为主节点使用，当整个集群使用不同数目的节点时，对于同一挖掘数据，假设最小支持度相同，图 5 为节点数目变化对挖掘的影响。

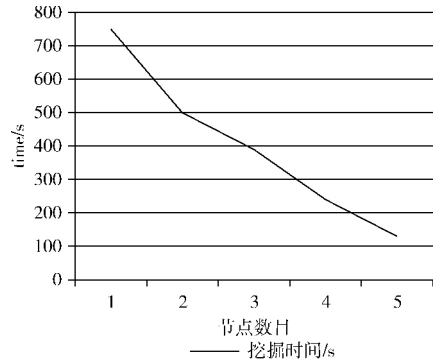


图 5 节点数目变化对挖掘的影响

图 5 中结果表明，集群中节点增多，挖掘时间减少，效率提高，这也体现了并行处理时分布式计算的优势，通

过灵活配置节点，可以轻松应对海量数据的处理。

## 4　实例分析

### 4.1　数据准备[10]

以我校信息工程学院 2006 年各专业学生大学四年学习成绩为研究对象，进行数据挖掘，找出课程间的隐含关系，为学生的学习进行数据决策支撑。表 2 是以学号为主键的学生四年学习成绩总表。

表 2　学生四年学习成绩总表

| 学号 | 高数 | 大外 1 | C 语言 | 数据结构 | 大计 | 大物 | … |
|---|---|---|---|---|---|---|---|
| 09060104 | 84 | 88 | 85 | 79 | 90 | 81 | … |
| 09060305 | 92 | 85 | 91 | 87 | 92 | 78 | … |
| 09060212 | 86 | 78 | 70 | 89 | 88 | 69 | … |
| 09060422 | 80 | 90 | 95 | 72 | 79 | 56 | … |
| 09060134 | 74 | 80 | 65 | 69 | 85 | 88 | … |
| … | … | … | … | … | … | … | … |
| 09060321 | 95 | 91 | 72 | 86 | 88 | 74 | … |

### 4.2　数据清理及转换

首先进行数据的筛选，在所有的学生数据对象中，存在空值的以及选修课程对应的数据项，都不予以考虑。接下来进行数据的转换，为了方便统计并且输入系统，用数字来代替不同的课程名称，并且划分规范的标准：按分数＞＝90、80-89、70-79、60-69、＜＝60 分别划分等级优、良、中、及格、不及格。

经过转换以后的数据见表 3。

表 3　数据转换

| 1 | 2 | 3 | 4 | 5 | 6 | … |
|---|---|---|---|---|---|---|
| 84 | 88 | 85 | 79 | 90 | 81 | … |
| 92 | 85 | 91 | 87 | 92 | 78 | … |
| 86 | 78 | 70 | 89 | 88 | 69 | … |
| 80 | 90 | 95 | 72 | 79 | 56 | … |
| 74 | 80 | 65 | 69 | 85 | 88 | … |
| … | … | … | … | … | … | … |
| 95 | 91 | 72 | 86 | 88 | 74 | … |

表 3 的成绩再按照“＞＝90”、“＞＝80”、“＞＝70”、“＞＝60”分别统计相应的课程数字编码输入系统，来对各专业的学生成绩数据进行关联规则，由于在 60 分以下学生数据不多，因此，为了统计方便，暂时不考虑不及格的学生数据。

### 4.3　关联规则统计

（1）计算机专业

表 4 统计出计算机专业学生成绩的关联规则。

表 4　关联规则（一）

| 1 | 关联规则 | 支持度 | 置信度 |
|---|---|---|---|
| ＞＝90 | 高等数学＝＞概率论与数理统计 | 0.21 | 0.48 |
| | 概率论与数理统计＝＞高等数学 | | 0.66 |
| ＞＝80 | C 语言＝＞数据结构 | 0.24 | 0.58 |
| | 数据结构＝＞C 语言 | | 0.85 |
| ＞＝70 | JAVA＝＞智能手机开发 | 0.55 | 0.69 |
| | 智能手机开发＝＞JAVA | | 0.86 |

由此可见，概率论与数理统计优秀的同学中高等数学优秀的可能性为 66%，而高等数学优秀的同学中概率论与数理统计优秀的可能性只为 48%。数据结构良好的同学中 85%的同学 C 语言良好，智能手机开发成绩中等的同学中 86%的人 JAVA 学得都很好。

（2）电子信息工程专业

表 5 统计出电子信息工程专业学生成绩的关联规则。

表 5　关联规则（二）

| 1 | 关联规则 | 支持度 | 置信度 |
|---|---|---|---|
| ＞＝90 | 模电实验＝＞数电实验 | 0.23 | 0.88 |
| | 数电实验＝＞模电实验 | | 0.66 |
| | 高等数学＝＞数电实验 | 0.21 | 0.78 |
| | 数电实验＝＞高等数学 | | 0.54 |
| ＞＝80 | 模电实验＝＞数电实验 | 0.34 | 0.91 |
| | 数电实验＝＞模电实验 | | 0.68 |
| | 高等数学＝＞数电实验 | 0.28 | 0.72 |
| | 数电实验＝＞高等数学 | | 0.56 |
| ＞＝70 | 模电实验＝＞数电实验 | 0.56 | 0.85 |
| | 数电实验＝＞模电实验 | | 0.78 |

由此可见，高等数学优秀的同学中数电实验优秀的可能性为 78%，模电实验好的同学数电实验也好，同时数电实验好的同学模电也好。

（3）物理师范专业

表 6 统计出物理师范专业学生成绩的关联规则。

表 6　关联规则（三）

| 1 | 关联规则 | 支持度 | 置信度 |
|---|---|---|---|
| ＞＝90 | 高等数学＝＞数电实验 | 0.21 | 0.87 |
| | 数电实验＝＞高等数学 | | 0.68 |
| ＞＝80 | 高等数学＝＞物理实验 | 0.34 | 0.82 |
| | 物理实验＝＞高等数学 | | 0.64 |
| ＞＝70 | 物理实验＝＞数电实验 | 0.78 | 0.92 |
| | 数电实验＝＞物理实验 | | 0.94 |

由此可见，高等数学优秀的同学中数电实验优秀的可能性为 87%，高等数学良好的同学中物理实验良好的可能性为 82%，物理实验好的同学数电实验也好，数电实验好的同学物理实验也同样好。

### 4.4　结果分析

根据实验结果，可以得出结论，高等数学、概率统计

课程是非常重要的基础课程，其自身相辅相成，对其它专业课程如数电、物理的学习起着至关重要的作用。专业课程中，先修课程例如C语言、JAVA、模电等，会对后续的数据结构、智能手机开发、数电等起着良好的铺垫作用，学好了先修课程才能更进一步地学好更多专业课程。工科同学实验课是必不可少的，其理论课程的学习非常重要，但实验课也并不完全依赖理论课，如高等数学、概率统计这样的基础课也很重要，经过统计发现高等数学成绩好的同学实验成绩也很好。最后发现，该院很多同学实验动手能力较强，应该进一步加强理论课的学习。

## 5 结束语

本文应用了目前大数据领域较流行的技术 Hadoop，对传统的 Apriori 算法进行了并行化的改进，通过进一步实验的验证，表明改进后算法在保证挖掘结果正确的前提下，执行效率更高，算法运行的稳定性更强，当挖掘数据量进一步增大时，可以利用分布式的方式增加集群中的节点，发挥分布式集群的最大优势进行数据挖掘。最后对本科计算机学院的学生数据进行分析，发现了课程间的隐含规则，对于学生重视课程、安排计划、调整作息时间具有指导性的意义，大数据技术在教育层面的应用将逐渐成熟、范围更广。

## 参考文献：


[1] Casado R，Younas M. Emerging trends and technologies in big data processing [J]. Concurrency & Computation Practice & Experience，2015，27 (8)：2078-2091.

[2] Oruganti S，Ding Q，Tabrizi N. Exploring Hadoop as a platform for distributed association rule mining [C] //The Fifth International Conference on Future Computational Technologies and Applications. Valencia：IEEE，2013：62-67.

[3] Zhou X，Huang Y. An improved parallel association rules algorithm based on MapReduce framework for big data [C] //International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Xiamen：IEEE，2014：284-288.

[4] Ezhilvathani A，Raja K. Implementation of parallel Apriori algorithm on Hadoop cluster [J]. International Journal of Computer Science & Mobile Computing，2013，2 (4)：513-516.

[5] ZHOU Guojun. Study on multi-keyword sorting method based on Hadoop [J]. Computer Engineering and Applications，2016，52 (17)：79-83 (in Chinese). [周国军. 基于 Hadoop 的多关键字排序方法研究 [J]. 计算机工程与应用，2016，52 (17)：79-83.]

[6] GAO Xianwei. Research and design of context aware recommendation system based on Hadoop [D]. Taiyuan：North University，2015：7-8 (in Chinese). [高献卫. 基于 Hadoop 的上下文感知推荐系统研究与设计 [D]. 太原：中北大学，2015：7-8.]

[7] Loughran S，Calero JMA，Farrell A，et al. Dynamic cloud deployment of a MapReduce architecture [J]. IEEE Internet Computing，2012，16 (6)：40-50.

[8] CHEN Wei. Construction method of education cloud storage system based on Hadoop platform [J]. Chinese Medical Education Technology，2015，29 (1)：29-33 (in Chinese). [陈伟. 基于 Hadoop 平台的教育云存储系统的构建方法 [J]. 中国医学教育技术，2015，29 (1)：29-33.]

[9] WEI Ling，WEI Yongjiang，GAO Changyuan. Improvement of Apriori algorithm based on Bigtable and MapReduce [J]. Computer Science，2015，42 (10)：208-210 (in Chinese). [魏玲，魏永江，高长元. 基于 Bigtable 与 MapReduce 的 Apriori 算法改进 [J]. 计算机科学，2015，42 (10)：208-210.]

[10] ZHANG Guohua. Application and practice of data mining in curriculum early warning of independent colleges [J]. Modern Electronic Technology，2016，39 (17)：136-139 (in Chinese). [张国华. 数据挖掘在独立学院课程预警中的应用与实践 [J]. 现代电子技术，2016，39 (17)：136-139.]