doi:10.3969/j.issn.1673-9833.2016.04.007

一种改进的基于Spark的用户行为分析方法的研究

阮得宝，李长云

（湖南工业大学 计算机与通信学院，湖南 株洲 412007）

摘　要：为解决大数据量情况下的网络用户行为分析的时效性、准确性，针对Apriori算法对数据库反复扫描和候选集过大的问题，提出了一种将压缩矩阵和事务权值引入的改进型Apriori算法，并将改进后的算法运用于云计算平台Spark。实验证明，改进后的算法的性能和效率都更高，在网络用户行为分析中具有优势。

关键词：Spark；Apriori；互联网；数据分析；网络用户行为分析

中图分类号：TP301.6　　**文献标志码**：A　　**文章编号**：1673-9833(2016)04-0032-04

On an Improved Spark-Based Method for the Analysis of User Behaviors

RUAN Debao，LI Changyun

（School of Computer and Communication，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：In view of the repeated scanning of the database and the potential massive candidate sets involved in the Apriori algorithm, an improved method, with the compressed matrix and the transaction value introduced in the process, is proposed to solve such problems as the timeliness and accuracy of the analysis of network user behaviors, with a further application of the improved algorithm to Spark, a cloud computing platform. The experimental results verify the better performance and higher efficiency of the proposed method, with evident advantages in the user behavior analysis.

Keywords：Spark；Apriori；Internet-work；data analysis；user behavior analysis

1　背景知识介绍

随着互联网的飞速发展，网络逐渐成为人们获取信息的最重要手段，网络数据流量也产生了巨大的增长。用户行为分析，是指在获得网站访问量基本数据的情况下，对有关数据进行统计、分析，从中发现用户访问网站的规律，并将这些规律与网络营销策略和推荐系统等相结合，从而发现目前网络活动中可能存在的问题，并为进一步优化用户体验和扩大服务提供商利益提供依据[1]。但是由于网络的开放性、动态性以及多样性等特点，用户在网上产生的数据量越大，用户行为分析的难度也越大。因此，在大数据量情况下对网络用户行为进行分析的需求越来越迫切。

用户行为分析的目的，是掌握用户的行为习惯和特点，进而根据用户的行为特点进行有针对性的网络信息推送；通过推送，用户获取需要信息的难度也大大降低。用户行为分析方式主要有以下几种方法：用户特征分析，是指找出用户的行为特征的方法；关联分析，是指寻找用户的两种或者几种行为习惯的联系、相关性或者因果关系；分类与预测，

收稿日期：2016-06-17

基金项目：国家自然科学基金资助项目（61350011，61379058，41362015），湖南省自然科学基金资助项目（14JJ2115，12JJ2036），湖南工业大学自然科学基金资助项目（2014HX16）

作者简介：阮得宝（1991-），男，安徽六安人，湖南工业大学硕士生，主要研究方向为大数据，E-mail：1040668038@qq.com

利用分类技术将用户归属于一个特定的类；异常分析，针对用户的不正常网络流量进行分析；Top*N*分析，在用户行为分析中，往往按照某一指标进行倒序或者正序排列，取前*N*项分析。

互联网发展的同时，网络用户行为相关的数据也在激增，传统的用户行为分析方法不足以支持如此巨大的数据量。因此，用户行为分析的方法必须运用海量数据运算[2]。在这样的情况下，海量数据的挖掘技术就至关重要。当前海量数据挖掘的办法主要有以下几种。

1）抽样，对数据进行抽样，在抽样数据的基础上建立数据挖掘模型；2）集成方法，划分数据集，并行创建分类器，然后集成处理；3）Map/reduce框架，基于云计算平台处理海量数据[3]。

针对云计算情形下的用户行为分析的算法主要有以下几种。

1）分类。分类是指将数据库中的数据按照种类和性质分别归类。2）回归分析。回归分析是指查找出几种变量之间的依赖关系，并用来分析变量里所包含的数据之间的规律。3）聚类分析。聚类分析是指根据规定的聚类变量，将数据库中的数据分成若干类。4）关联规则。关联规则是指数据对象之间的依赖关系，目的就是从发现支持度大于给定值的规则。5）神经网络方法。神经网络方法模拟人的直观思维方式，将信息分布式存储以及并行协同处理，特点是非线性映射能力及高度的并行性，神经网络方法在逻辑推理中写成串行的指令，让计算机运行。6）Web数据挖掘。Web数据挖掘应用于Web环境，它从大量的Web文档数据中发现隐藏在数据中的规律，通过对这些数据的挖掘，可以得到仅通过文字检索无法获得的信息。

基于云计算平台的用户行为分析中，较为重要、应用较广泛的算法是关联规则算法，其流程图如图1所示。

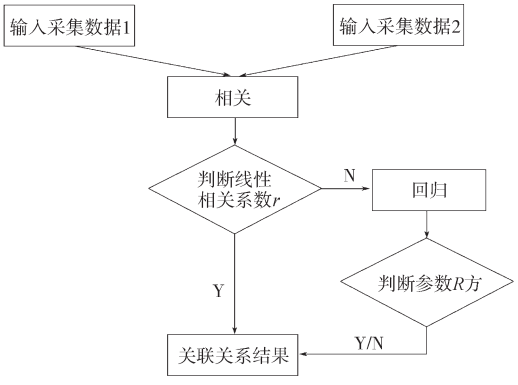

图1　关联分析流程图

Fig. 1　Flow chart of the correlation analysis

关联分析中最典型的算法就是Apriori算法。Apriori算法的核心思想是基于两阶段频集思想的挖掘算法，它的优点是简单、容易理解和数据要求低。但是，传统的Apriori算法存在下面2个缺点：1）根据Apriori算法的定义，Apriori算法会产生大量的频繁集；2）在数据库规模巨大的情况下，算法会重复扫描事务数据库，这种反复扫描会增加I/O的负载，并且随着数据库规模的增加，I/O负载会呈现指数式的增加。

近几年，随着Hadoop的使用不断广泛和完善，大量研究人员也都致力于将传统的数据挖掘算法采用并行编程框架进行分布式并行化处理，以提高数据挖掘的效率[4]。许多学者也针对Apriori算法的不足，结合Hadoop在大集群中所展现的优势，提出了一些基于Hadoop的改进型Apriori算法。其不足之处大多是求解局部的频繁项集时，没有剪切操作，导致生成的候选项集过大。随着Spark的异军突起，Hadoop有逐渐被取代的趋势。现阶段，针对Spark的运用，主要集中在数据挖掘方面，鲜有将Spark运用于用户行为分析的研究。本文提出的基于Spark的改进型Apriori算法更能适用于大数据挖掘，从而对用户行为进行可靠的分析。

2　基于Spark的Apriori算法设计

针对Apriori算法效率欠佳的问题，许多研究人员在Apriori原始算法的基础上对Apriori算法进行了大量地改进。改进的Apriori算法所使用的主要技术有：哈希技术、事务压缩技术、分区技术和采样技术等等。而通过这些技术改进后的Apriori算法都存在着候选集过大或者牺牲了原算法计算结果准确性的问题[5]。

在大数据情况下，影响Apriori算法效率最大的问题主要是对数据库的反复扫描。为了解决这个问题，课题组在使用矩阵形式来存储数据库的基础上，提出利用向量计算的方法计算支持度，并且对相同事务压缩以减小矩阵的大小，优化数据结构，提高算法的效率。算法首先对交易事务的数据库进行扫描，然后将数据库转化成一个布尔矩阵，转化的同时对事务数据进行压缩，在生成频繁集时，得到项集的支持度计数，若大于最小支持度计数，保留项集，反之则舍弃。计算过程中对数据矩阵进行删减，并反复迭代上述过程，避免了原算法在计算候选集的支持度时扫描全部数据库的不足。

Spark是UC Berkeley AMP Lab所开源的类Hadoop

Map/reduce 的通用并行框架，拥有 Hadoop MapReduce 所具有的优点。Spark 与 Map/reduce 的区别是，它的 Job 中间输出和结果可以保存在内存中，没有读写 HDFS 的需求[6]。因此，Spark 为迭代式数据处理提供了更好的支持。Spark 的生态系统如图 2 所示。Spark 的核心和精华是它的弹性分布式数据集（resilient distributed dataset，RDD），RDD 是只读的纪录分区的集合，能够在内存中加载，方便再次使用。RDD 可以分布在多个节点上，可以进行并行处理[7]。

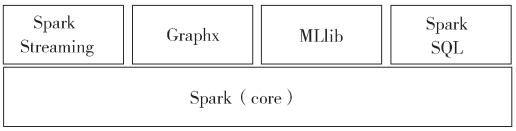

图 2 Spark 的生态系统图

Fig. 2 Spark ecosystem diagram

基于 Spark 的改进型 Apriori 算法使用 Map/reduce 思想实现频繁的计算。基于 Spark 的改进 Apriori 算法流程图如图 3 所示。

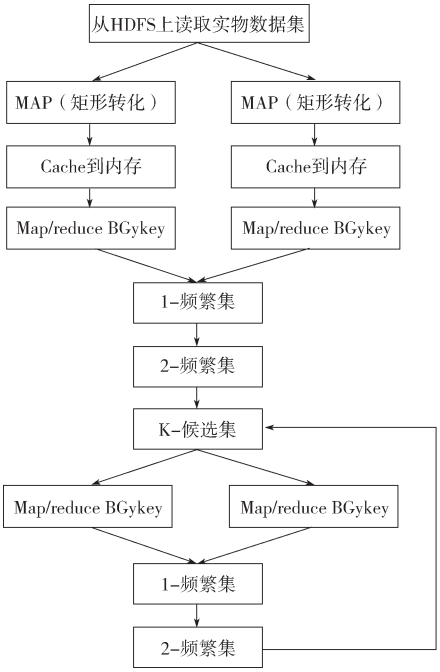

图 3 基于 Spark 的改进 Apriori 算法流程图

Fig. 3 Flow chart of the improved Spark-based Apriori algorithm

在设计方面，由于 Spark 运行在 Mesos 平台，所以采用分布式管理系统作为原始数据存放的存放系统。在 Spark 实现算法时，首先将原始的交易数据存放在 HDFS 中，随后读取 HDFS 上的交易事务数据，并将其转化为压缩矩阵，根据转化后的矩阵数创建 RDD。其次，通过 Map 操作计算候选集的局部支持度计数，通过 Reduce 计算候选集的全局支持度计数。因为转化后的数据是缓存到本地随机存取存储器 RAM 中，所以每一个 Map 操作的过程中，都是直接读取候选项集中项的行向量数据，不需要到分布式文件系统上重写。数据集的划分和任务分配都在 Spark 中由系统自动完成的[8]。为了验证改进算法的性能提升程度，课题组进行了 2 组实验，分别检验两种算法在常规和云计算平台下的表现。

3 实验结果分析

Apriori 算法性能提升，是指算法在分析处理大数据环境下的用户行为数据时，其运算速度得到了提升。课题组设计 2 组实验来验证改进算法的性能。实验一在不同规模的数据环境下，检验 2 种算法的性能。实验二则是检测在云计算平台下，针对不同节点时，算法性能。实验环境配置如下表 1 所示。

表 1 实验环境配置表

Table 1 Configuration table of the experimental environment

项目	配置
Spark 版本	Spark-1.2.0
Hadoop 版本	Hadoop-2.5.0
操作系统	Ubuntu-12.04
Master	4 核，16 G 内存，500 G 硬盘，数量 1
Woker	4 核，16 G 内存，80 G 硬盘，数量 5

实验采用的数据是从数据堂获得，数据内容为社交资源共享站点用户行为数据集[9]，数据大小总计 652.36 MB。实验一采用 3 种数据规模，对原算法和改进算法的性能进行对比，如表 2 所示。图 4 用折线图方式对比 2 种算法的运行时间。

表 2 实验 1 算法测试结果

Table 2 Algorithm test results of experiment 1 ms

算法	数据大小		
	50 MB	100 MB	500 MB
Apriori	331	729	2 252
改进 Apriori	116	428	1 672

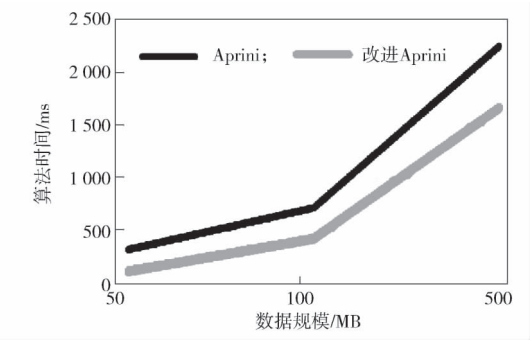

图 4 不同数据规模下 2 种算法性能的比较

Fig. 4 Comparison between the performance of two algorithms with different data scales

通过表 2 和图 4 可知，改进 Apriori 算法效率在执行时间上得到了很大的提升。但当数据量增大时，提

升的效率有所降低。

实验二使用不同的节点的集群，测试处理相同大小数据时，原算法和运行于云计算平台Spark的改进算法的性能。原算法无法运用在多节点，因此仅有1组数据。实验结果如表3所示。图5用折线图的方式直观地反映在不同节点数量情况下，改进算法的性能。

表3　实验2算法测试结果

Table 3　Algorithm test results of experiment 2　　ms

算法	节点数		
	1	2	3
Apriori	2 252		
改进Apriori	1 672	542	387

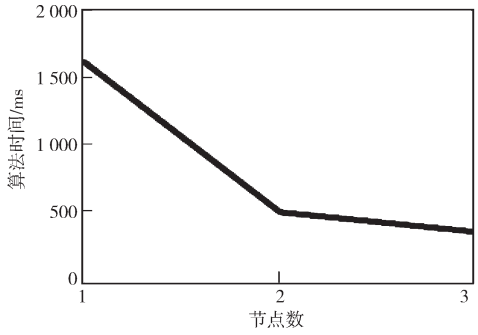

图5　不同节点下算法的性能

Fig. 5　Improved performance of algorithms with different nodes

通过表3和图5可知，随着节点数量的上升，在云计算平台上使用的改进Apriori算法的效率越来越高。总的来说，效率的提升和节点数成正比。

综合以上的实验结果可以得出结论，在云计算平台上使用的改进Apriori算法在性能和效率方面提升显著，符合预期。

4　结语

本文为解决大数据量情况下的网络用户行为分析的时效性、准确性，针对Apriori算法对数据库反复扫描和候选集过大的问题，在使用矩阵形式来存储数据库的基础上，提出了利用向量计算的方法计算支持度，并且对相同事务压缩以减小矩阵的大小，优化数据结构，提高算法的效率，并将改进后的算法运用于云计算平台Spark。实验结果表明，课题组提出的改进Apriori算法的性能有明显的提升。并且改进后的算法运用于云计算平台Spark时，性能也有进一步显著的提升。

参考文献：

[1] 吕桃霞，刘培玉. 一种基于矩阵的强关联规则生成算法[J]. 计算机应用研究，2011，28(4)：1301-1303.
LÜ Taoxia，LIU Peiyu. Algorithm for Generaring Strong Association Rules Based on Matrix[J]. Application Research of Computers，2011，28(4)：1301-1303.

[2] 刘宗成，张忠林，田苗凤. 基于关联规则的网络行为分析[J]. 电子科技，2015，28(9)：16-18.
LIU Zhongcheng，ZHANG Zhonglin，TIAN Miaofeng. Analysis of Network Behaviors Based on Association Rules[J]. Electronic Science and Technology，2015，28(9)：16-18.

[3] Apache Spark. The Apache Software Foundation[EB/OL]. [2015-09-21]. http://spark.apache.org.

[4] 吴岳忠，周训志. 面向Hadoop的云计算核心技术分析[J]. 湖南工业大学学报，2013，27(1)：77-80.
Wu Yuezhong，Zhou Xunzhi. The Core Technology of Hadoop-Oriented Cloud Computing[J]. Journal of Hunan University of Technology，2013，27(1)：77-80.

[5] 陶彩霞，谢晓军，陈　康，等. 基于云计算的移动互联网大数据用户行为分析引擎设计[J]. 电信科学，2013，29(3)：27-31.
TAO Caixia，XIE Xiaojun，CHEN Kang，et al. Design of Mobile Internet Big Data User Behavior Analysis Engine Based on Cloud Computing[J].Telecommunications Science，2013，29(3)：27-31.

[6] 郑凤飞，黄文培，贾明正. 基于Spark的矩阵分解推荐算法[J]. 计算机应用，2015，35(10)：21-23.
ZHENG Fengfei，HUANG Wenpei，JIA Mingzheng. Matrix Factorization Recommendation Algorithm Based on Spark[J]. Journal of Computer Applications，2015，35(10)：21-23.

[7] 宋文惠，高建瓴. 基于矩阵的Apriori算法改进[J]. 计算机技术与发展，2016，26(6)：80-83.
SONG Wenhui，GAO Jianling. The Inproved Apriori Algorithm Based on Matrix[J]. Computer Technology and Development，2016，26(6)：80-83.

[8] 李仕琼. 数据挖掘中关联规则挖掘算法的分析研究[J]. 电子技术与软件工程，2015(4)：200.
LI Shiqiong. Analysis of Algorithms Data Mining Association Rules Mining[J]. Electronic Technology & Software Engineering，2015(4)：200.

[9] YAO Junjie，CUI Bin，HAN Qiaosha，et al. Modeling User Expertise in Folksonomies by Fusing Multi-Type Features[C]//15th International Conference on Database Systems for Advanced Applications(DASFAA 2011). HongKong：Springer，2011：53-67.

（责任编辑：申　剑）