基于关联规则 Apriori 算法的物联网海量数据挖掘系统研究

周小云

(周口师范学院计算机科学与技术学院 河南，周口 466001)

Research on IOT massive data mining system based on Apriori algorithm of association rules

Zhouxiaoyun

(School of computer science and technology, Zhoukou Normal University, Zhoukou 466001, Henan)

Abstract: with the rapid progress of Internet of things technology, mining a large number of data generated by Internet of things has become a hot topic in the research field. How to efficiently analyze, process, store and mine these data to extract valuable information and assist the business decision-making related to the Internet of things is the key problem to be solved. Based on the in-depth analysis of the data characteristics of the Internet of things, this paper proposes a data mining model based on association rule algorithm, and integrates the model into the data mining system architecture. Through experimental verification, the model shows good performance, can effectively mine a large number of data, and provide strong support for business decision-making.

Keywords: Apriori algorithm of association rules; Internet of things; data mining

摘要：随着物联网技术的飞速进步，对物联网产生的大量数据进行挖掘已成为研究领域中的热点议题。如何高效地对这些数据进行分析、处理、存储和挖掘，以提取有价值的信息并辅助物联网相关的商业决策，是当前亟待解决的关键问题。本研究在深入分析物联网数据特性基础上，提出了一种基于关联规则算法的数据挖掘模型，并将该模型集成至数据挖掘系统架构中。通过实验验证，该模型展现出了良好的性能，能够有效地挖掘大量数据，并为商业决策提供有力支持。

关键词：关联规则 Apriori 算法；物联网；数据挖掘

收稿日期：2025 年 10 月 9 日 **中图分类号：**TP391.1

通讯作者：*周小云，周口师范学院计算机科学与技术学院

物联网在互联网基础上，技术与功能实现了持续升级，达成了用户对信息的感知、收集与传感。然而，借助物联网开展信息交换和通信时，会产生海量数据(像 RFID 数据流、传感器网络数据等)，这些数据的持续增加提升了用户从中获取有用信息的难度[1]。为提升物联网的数据处理能力，相关研究人员融合应用云计算、数据挖掘技术，搭建百万计算机集群的云模式，以完善物联网的弹性可扩展技术、分布式计算技术和存储机制，进而强化物联网的可信计算功能，有助于物联网在面对海量业务数据时能够迅速进行分析、处理、存储、挖掘，从而实现有价值信息的快速提取，服务于物联网商业决策。物联网海量数据挖掘依旧是研究热点，本文将结合数据挖掘技术里的关联规则 Apriori 算法、云计算技术、PML 来探索与设计物联网海量数据挖掘系统[2]。

1 物联网海量数据挖掘

1.1 物联网海量数据的特点

在物联网的应用进程中，会产生海量的数据。经研究可知，这些海量数据具备 4 大特点。①数据量庞大。传感设备是每个物联网系统的基础设备，数量多达成千上万，其作用是把采集到的数据传输至数据中心。为实现对象的状态跟踪、数据统计分析和数据挖掘，数据中心既要保存历史数据，又要接收当前采集的数据，大量数据增加了物联网数据中心的负担。②数据类型繁杂。因为物联网系统的监控对象种类繁多，不同种类监控对象采集的信息也不一样，所以物联网系统的数据类型（如文本、图像、视频等）十分复杂。③数据的异构性[3]。GPS 传感终端、RFID 传感终端等多种传感终端构成了物联网系统，不同的传感终端使得终端采集的数据在语义、格式上各不相同，导致物联网系统数据具有异构性，这在一定程度上增加了数据存储和挖掘的难度。④数据的高度动态性。物联网系统中有大量的传感终端和传感节点，不同的传感终端在系统的每一时刻可能被添加或被移除。因此，物联网系统数据库中可能会增加传感节点存储采集的数据，一旦传感节点被移除，其存储在系统数据库中的数据将不再保留记录。所以，传感节点的频繁动态变化让物联网系统中的数据呈现出高度动态性的特点[4]。

1.2 物联网海量数据挖掘

RFID 信息数据是研究物联网海量数据挖掘问题的主要对象，结合数据挖掘技术可以从该对象中挖掘出潜在且有价值的信息。RFID 传感器能够采集到 EPC（标签的标识码）、Location（阅读器读取标签的地点）、Time（阅读器读取标签的时间）这 3 个原始数据。这些数据的特征主要表现为海量性、分布式、时间与空间性、异构性、动态性、节点资源有限性，因此，精确挖掘物联网海量数据的难度极大[5]。在实际领域中，RFID 数据流分析、频繁与序列模式分析、分类与聚类的路径分析等是 RFID 信息数据挖掘的主要内容，这些数据的挖掘对物联网商业决策具有重要意义。

2 物联网海量数据挖掘系统的数据处理模式

2.1 物联网海量数据挖掘系统处理海量数据的流程

处理物联网海量数据挖掘中的 RFID 动态异构数据，需要基于云计算技术和数据挖掘技术，以 Hadoop 为平台，利用 Map/Reduce 模式来实现数据挖掘处理。具体操作流程如下：①对物联网中的 RFID 数据进行过滤、转换、合并，以 PML 文件的形式保存在分布式系统 HDFS 中。为解决高效存储、处理和节点失效的问题，可以采用副本策略，将 PML 文件的 2 – 3 个副本保存在同一机构的不同节点上，或者保存在不同机构的某一节点上。②主控程序 Master 在执行任务时主要负责创建和管理控制任务，空闲状态的 Worker 会收到相关分配任务，并配合 Map/Reduce 进行操作处理，之后由 Master 归并最终结果并向用户反馈[6]。

2.2 计算和存储的整合及迁移

由于系统采用分布式数据存储方式，所以能够实现计算与存储的整合和迁移，这也是基于云计算、关联规则 Apriori 的物联网海量数据挖掘系统的一大特点。系统计算和存储的整合及迁移处理过程需要借助 Map/Reduce 模式，具体实施策略是在本地计算机上操作[7]。因为 Map 在每个节点上的操作相互独立，不存在数据传输，只有在 Reduce 过程中需要将计算结果传送给 Master，有利于实现计算和数据的同步密集以及计算向存储的迁移，从而大大加快了数据传输时间。同时，系统还应用了 PML 文件副本策略，当出现节点失效时，DataNode 节点会有一个副本节点供 Master 使用，该副本节点会实现计算迁移（此过程中数据不会在 DataNode 节点间传递）并重新开始数据处理，这样就不必重启全部工作，大大提高了数据传输效率。

具体的 Map/Reduce 操作过程如下：

①运用 Map/Reduce 思想，根据参数将输入文件分割成大小在 16 – 64M 范围的 M 块；②执行程序主要包括主控程序 Master 和分工作机 Worker，其中 Map 操作有 M 个，Reduce 操作有 R 个，空闲的 Worker 会接收 Master 分配的 Map 或 Reduce 处理任务；③ Worker 在处理 Map 任务时会读取处理数据，然后将 <key，value> 传递给 Map 函数并产生中间结果，将其缓存在内存中，定时将缓存的中间结果传送到本地硬盘，通过分区函数将其划分为 R 个块区，把本地硬盘接收数据的位置信息通过 Master 传送给 Reduce 函数；④ Reduce Worker 根据 Master 传送的文件信息，通过远程读取方式

找到对应的本地文件，对文件中的中间 key 进行有序排列，再通过远程向具体执行的 Reduce 发送信息；⑤ Reduce Worker 将排序后的中间数据的 key 和相应的中间结果集传送给 Reduce 函数，并以最终输出文件编写最后的结果；⑥完成所有的 Map 和 Reduce 任务后，MapReduce 返回用户程序的调用点，并由 Master 激活用户程序[8]。

3 物联网海量数据挖掘系统的设计

如图 1 所示，这是基于关联规则 Apriori 算法与云计算的物联网海量数据挖掘系统的基本结构。

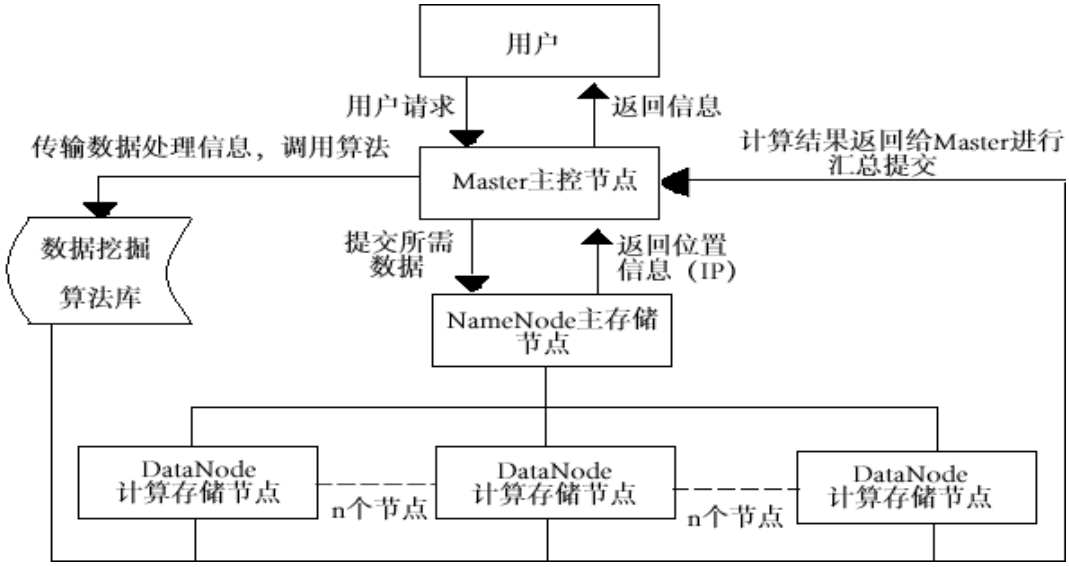

图 1 系统基本结构

本研究提出的系统架构主要由三个层次构成：数据存储层、数据挖掘算法层以及挖掘任务处理层。在该架构中，主控节点（Master）扮演着至关重要的角色，负责与用户交互、调度以及管理整个系统的节点任务[9]。系统中采用 Map/Reduce 化处理的数据挖掘算法被部署于特定节点，以提升数据挖掘过程的效率。在 HDFS 分布式存储系统中，系统由单一 NameNode 主节点和多个 DataNode 构成。NameNode 主节点负责处理用户请求，并向用户返回存储数据的 DataNode 的 IP 地址，同时向其他 DataNode 发送数据副本的通知。

在数据挖掘算法层，本研究将数据挖掘领域中常用的算法进行了 Map/Reduce 化处理。以分布式并行关联规则算法为例，该算法是通过将 Apriori 算法 Map/Reduce 化得到的。这些算法集成于系统数据挖掘算法层的算法节点中。在实际应用中，通过云计算平台的辅助，利用 Master 主控节点进行控制与管理，根据用户需求向相关节点分发算法以执行计算任务[10]。

在挖掘任务处理层，该层相当于系统的任务调度层，是系统的核心。Master 节点负责调度系统中所有的挖掘器。具体的挖掘任务处理流程包括：首先，Master 节点会识别并记录空闲的 DataNode 节点，将其纳入空闲节点列表；其次，用户请求由 Master 节点接收，并获取 DataNode 中各个数据块的存储信息以及挖掘调用算法；再次，Master 节点向算法存储节点申请所需的挖掘算法，并由算法存储节点将算法传送给 DataNode 节点（原始数据）；最后，在 HDFS 服务器中根据计算任务启动工作，工作完成后将结果传送给 Master 节点。Master 节点汇总后生成最终结果并反馈给用户。由于该过程无需进行数据重组与传送，系统中每个节点的计算和存储文件传输效率得到显著提升[11]。

4 基于关联规则 Apriori 的数据挖掘算法

尽管数据挖掘算法分类众多，但在物联网数据挖掘中最有效的还是关联规则的 Apriori 算法[7]。Apriori 算法运用逐层搜索迭代方式来通过 K 项集进行（K+1）项集的探索，首先需对数据集进行一次扫描，进而生成频繁 1– 项集 L_1，之后利用 L_1 进行频繁项集 L_2 的探索，以不断迭代

的方式持续到频繁项集为空集。由于频繁项集具有任一子集都为频繁项集的特性来压缩处理搜索空间，以此加快频繁项集的生成效率[12]。在经历了第 K 次循环搜索后，数据挖掘的具体过程有：①操作 JOIN（连接），令 L_{K-1} 产生候选集 C_K 并进行连接操作；②按照 Apriori 性质来完成支持度统计与剪枝的操作，令 C_K 产生频繁集 L_K。这种算法的不足之处是需要多次扫描数据库才可探索出所有的频繁项集，显然具有海量数据的物联网应用并不适合这一算法，多次扫描会耗损大量内存及时间[13]。因此，本文借鉴云计算平台的分布式并行计算性质，将该性质移植在 Apriori 算法上，建立 Hadoop 架构以存储扫描数据库查找频繁项集所获得的并联规则，扫描处理将在各个 DataNode 节点中并行操作，由此获得各计算节点上的局部频繁项集。之后，利用 Master 将实际的全局的支持度、频繁项集统计与确定出来，以此来节省系统的时间与内存消耗，实现数据挖掘效率的大大提高[14]。

同时，还需对 Apriori 算法进行 Map/Reduce 化，具体处理流程有：①用户请求挖掘服务，并将关联规则需要的最小支持度、置信度由用户来设置；②接收到请求的 Master 需向 NameNode 申请相关的 PML 文件，对空闲节点列表进行访问，分配任务给空闲的 DataNode，将各个 DataNode 所需的存储算法节点的算法进行调度与并行处理；③将每个 DataNode 利用 Map 函数进行 <key，value> 对映射与新键值对的处理，生成一个局部候选频繁 K 项集，用 C^n_K 来表示，每一 C^n_K 的支持度用 1 表示；④利用 Reduce 函数进行调用计算，累加每个 DataNode 节点上相同的候选项集的支持度，以生成一个实际的支持度，对比用户申请时设置的最小支持度，以产生局部频繁 K 项集的集合，用 L^n_K 表示；⑤合并所有的处理结果，以生产全局频繁 K 项集 L_K。如图 2 所示，其是 Apriori 算法 Map/Reduce 化的具体实现流程[15]。

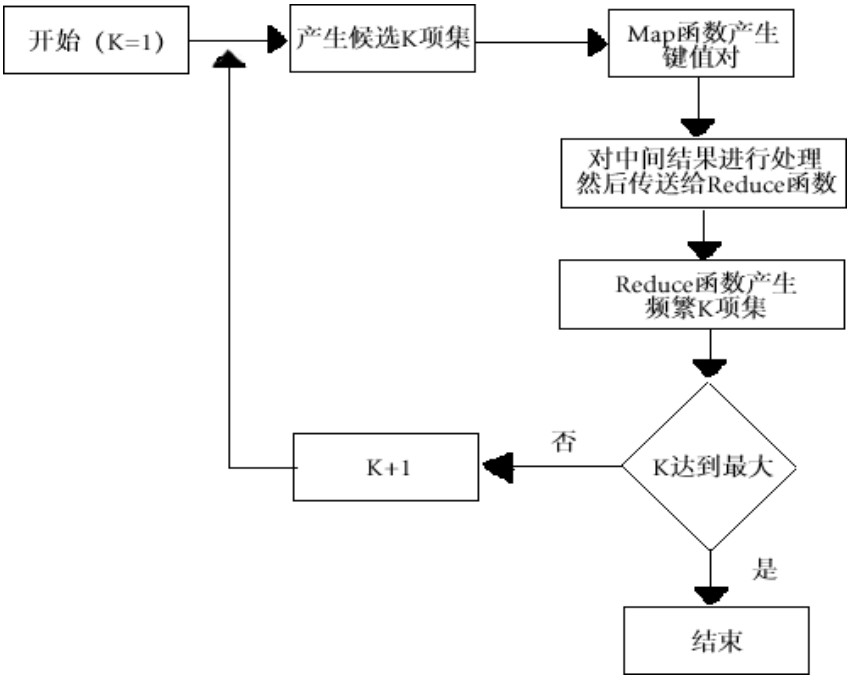

图 2 M ap/ Reduce 化的 Apriori 挖掘算法实现流程

5 结语

鉴于数据挖掘在物联网应用中的重要性，以及物联网海量数据挖掘面临的问题，本文依托关联规则 Apriori 算法和云计算技术，开展物联网海量数据挖掘系统的研究与设计。同时，结合物联网数据特性，在数据挖掘技术和云计算技术支撑下，提出物联网海量数据挖掘算法，完成物联网海量数据挖掘系统设计，进而有助于有效解决物联网海量数据挖掘问题。

参考文献：

[1] 张明, 李强. 基于改进 Apriori 算法的物联网设备数据关联规则挖掘方法[J]. 计算机研究与发展,2023,60(5):1123-1134.

[2] 王静, 刘洋. 面向物联网大数据的并行化 Apriori 算法优化研究[J]. 计算机应用研究,2022,39(8):2456-2462.

[3] 陈晨, 赵芳. 基于 Apriori 算法和云计算的物联网数据挖掘系统设计[J]. 计算机科学,2021,48(10):328-335.

[4] 孙丽华, 吴晓峰. 改进 Apriori 算法在智慧城市物联网数据分析中的应用[J]. 通信学报,2020,41(12):156-165.

[5] 杨雪, 周涛. 基于 MapReduce 和 Apriori 的物联网大数据挖掘方法[J]. 计算机工程与应用,2022,58(15):102-110.

[6] 刘佳, 胡斌. 物联网环境下基于 Apriori 算法的异常数据关联分析[J]. 电子学报,2021,49(7):1421-1429.

[7] 黄敏, 林峰. 改进加权 Apriori 算法在工业物联网数据挖掘中的应用[J]. 计算机集成制造系统,2020,26(5):1265-1274.

[8] 徐阳, 郑洁. 基于 Spark 框架的 Apriori 算法并行化优化研究[J]. 计算机应用,2023,43(4):1145-1153.

[9] 李娜, 王磊. 面向智能家居的 Apriori 算法改进与数据挖掘研究[J]. 计算机工程,2021,47(9):154-162.

[10] 赵鑫, 马红梅. 基于 Apriori 算法的农业物联网数据关联规则挖掘[J]. 计算机应用与软件,2020,37(6):315-322.

[11] 吴迪, 周明. 物联网环境下基于 Apriori 和聚类分析的数据挖掘方法[J]. 计算机科学,2022,49(5):298-306.

[12] 刘芳, 陈刚. 改进 Apriori 算法在车联网大数据分析中的应用[J]. 计算机工程与设计,2021,42(8):2214-2220.

[13] 孙伟, 张丽. 基于 Hadoop 平台的 Apriori 算法并行化实现[J]. 计算机应用研究,2019,36(12):3675-3682.

[14] 胡静, 杨帆. 面向医疗物联网的改进 Apriori 数据挖掘算法[J]. 计算机工程与科学,2020,42(10):1842-1850.

[15] 郑阳, 刘伟. 基于 Apriori 算法的智能电网数据关联规则挖掘[J]. 计算机研究与发展,2023,60(3):678-688.

作者简介:周小云(1985-10), 女, 汉族, 河南周口人, 硕士, 周口师范学院计算机科学与技术学院讲师, 研究方向: 数据研究。