# 基于云计算平台的物联网数据挖掘研究

张毅，崔晓燕

（北京邮电大学自动化学院，北京 100876）

**摘 要：**随着社会的不断发展，信息产业已经逐渐成为了国民经济发展的重要支柱，而物联网作为新一代信息技术的重要组成部分成为推动人类文明向智能化方向发展的关键技术。物联网中的数据挖掘是物联网技术中重要的一环，是未来物联网应用数量大规模增长后对物联网产业的强力补充，本文分析了物联网数据的特点以及物联网数据挖掘存在的困难，以及云计算的出现为物联网数据挖掘提供了重要思路，文中论断云计算为物联网提供了最具计算力和 存储力的平台，并创新性的提出物联网云的概念。另外，在对平台可行性及性能进行分析的过程中，本文提出了数据转换器、开放平台接口等思路，使整个平台有更好的扩展性，方便第三方开发和测试。目前，物联网应用的整体生态系统面临很多挑战，产业链中的不同人群也面临着不同问题，本文也给物联网中这些问题的解决提供了很好的思路。

**关键字：**物联网；云计算；数据挖掘；分布式

**中图分类号：**TP319 **文献标识码：**A DOI：10.3969/j.issn.1003-6970.2014.1.034

**本文著录格式：**[1] 张毅，崔晓燕 . 基于云计算平台的物联网数据挖掘研究 [J]. 软件 ,2014,35(1):108-111

## The Data Mining Of IOT Based ON Cloud Computing

ZHANG Yi, CUI Xiao-yan

*(Automation School, Beijing University of Posts and Telecommunications, Beijing* 100876*)*

**【Abstract】**With the continuous development of society, the information industry has gradually become an important pillar of the national economy. As an important part of the new generation of information technology, the Internet of things promotes the development of human civilization and makes the world more intelligent. The data mining of IOT is an important part of IOT technologies and a strong complement to the IOT industry when more and more applications appear in the future. This paper analyzes the difficulties of data mining in IOT because of the data features of IOT, and cloud computing provides an important idea for data mining of IOT. The paper also concludes that cloud computing offers IOT the most powerful computing and storage capacity. In addition, the concept of IOT cloud is described in chapter two.In addition, during the analysis of the platform, this paper made a research of data converters and open platform interface, which makes the platform more scalability and easy to develop or test for the third party. At present, the applications ecosystem of IOT faces many challenges, and so does the IOT industry chain. This paper gives good ideas to solve these difficulties of IOT.

**【Key words】**internet of things; cloud computing; data mining; hadoop, distributed

## 0 引言

原 IBM 首席执行官彭明盛先生在 2010 年提出“智慧地球”[1] 的概念之后，物联网技术从研究阶段逐渐开始延伸到现实生活中，越来越多的物联网应用出现在人们的视野里，越来越多的公司加入到物联网发展的浪潮中来，物联网技术和产品都得到快速蓬勃发展，毫无疑问物联网将会成为互联网之后的又一个新的信息化热点。

这样也对目前物联网技术提出了更高的要求，物联网技术需要更深入地研究。正如我们在论文中反复提到的，物联网的数据有其自身的特征，如海量性、异构性等。尽管物联网数据种种特点让物联网的数据挖掘面临很多困难，但是物联网数据挖掘仍旧是物联网技术未来必须解决的问题。云计算的出现给物联网数据挖掘提供了很好的思路，我国在多个行业正在实施基于物联网的云计算平台，云计算让物联网的发展具备了更加强大的 IT 基础支撑能力 , 以及数据挖掘分析能力和平台开放扩展能力 [2]。所以说，物联网产业的发展 , 不仅自身发展潜力巨大，而且物联网和云计算的融合发展对社会经济产生更大的影响。

本文是基于云计算平台搭建的物联网数据挖掘研究，主要完成物联网数据在云计算平台上的优化存储和处理，云服务层主要面向数据挖掘，并根据物联网数据挖掘的特殊性改进数据挖掘算法，最终构建出一个面向物联网的数据挖掘平台。

## 1 物联网与云计算

### 1.1 物联网的概念

物联网（IOT）是下一代网络，包含上万亿节点来代表各种对象，从无所不在的小型传感器设备，掌上的到大型网络的服务器和超级计算机集群。它是继计算机和网络技术之后的又一场科技革命。它不仅包括了计算机技术和通信技术（如传感器网络，移动通信技术，RFID 技术，GPS，IPV6 等），同时还代表了下一代网络的发展方向。S. Haller 等人提出了如下的定义：“它是这样的一个世界，物理对象可以无缝集成到信息网络，并且可

以成为业务流程的积极参与者。服务可以在网络中影响到这些‘智能对象’，找到他们的国家以及与他们向关联的任何问题，并能考虑到安全和隐私问题[3]。”

物联网一般具有三个特征：首先是全面感知，主要表现在通过现有的一些技术，如电子标签等获得基本信息；其次是可靠的传递，这主要表现在信息的出书方面，包括了有线网络或无线网络，如传感器网络和其他通信网络（移动通信网、互联网等）将获取的物体信息可靠地传递出去；三是智能处理，物联网需要结合云计算、模糊识别等技术来处理多种来源的海量异构数据，同时要保证效率，有效的整合共享信息，达到真正对物体的智能控制。根据以上这三个基本特征分析一个典型的物联网应用至少应包含三个部分：（1）传感器、RFID、二维码等电子元器件；（2）数据处理中心，主要用于大量节点产生数据的存储和处理；（3）有线或无线的传输通道，例如 3G/4G、光纤等。

### 1.2 云计算技术

#### 1.2.1 云计算的概念

云计算 (CloudComputing), 是一种基于互联网的崭新的计算方式，通过互联网上异构、自治的服务为用户提供按需即取的计算。由于资源是在互联网上，而互联网常以一个云状图案来表示，因此可以形象地类比为云，“云”同时也是对 IT 底层基础设施的一种抽象概念，它是一种通过 Internet 以服务的方式提供动态可伸缩的虚拟化的资源的计算模式。

#### 1.2.2 Hadoop 概述

Hadoop[4] 可以被概括为由 Apache 软件基金会开发的一个分布式系统基础的架构。目前，Hadoop 被很多研究机构用来作为云计算的基础开发平台，它可以在用户不了解分布式底层细节的情况下进行分布式程序开发，由于 Hadoop 平台是开源的并且通过集群的优势提供了高速运算能力和强大的存储能力，因此被看做未来可以像 Linux 系统一样影响 IT 产业。Hadoop 是以分布式文件系统 HDFS 和 MapReduce 为核心，它提供了系统底层细节透明的基础架构，用户可以获得很好的分布式计算和分布式存储编程环境。HDFS 具有高容错性、高伸缩性等优点，使得用户可以不仅在服务器上部署 Hadoop，同时在低廉的硬件上也可以部署，形成分布式文件系统。MapReduce 分布式编程模型允许用户在不了解分布式系统底层实现细节的情况下开发并行应用程序，采用 MapReduce 来整合分布式文件系统上的数据，可保证分析和处理的高效性。用户可以利用 Hadoop 轻松地组织计算机资源，进而搭建自己的分布式计算云平台，并且可以充分利用集群的计算和存储能力，完成海量数据的处理。

## 2 数据挖掘技术

### 2.1 数据挖掘定义

数据挖掘的历史虽然不长，但是自上世纪九十年代以来，人们对其重视程度越来越大，由于数据挖掘属于一个交叉学科，不同的领域的人对其理解存在着不一样的地方，因此目前还没有一个统一的定义，不同的人根据自己的研究内容和应用对象提出了不同的定义： SAS 研究所认为数据挖掘是“在大量相关数据基础之上进行数据探索和简历相关模型的先进方法”； Bhavani 认为数据挖掘是“使用多种不同的技术，在大量的数据中发现有意义的新关系、模式和趋势的过程”；韩家炜等人认为“数据挖掘是在大型数据库中寻找有意义、有价值信息的过程”[5]。

大多数研究人员比较赞同韩家炜等人对数据挖掘的定义。这个定义主要包含几层含义：首先，数据来源必须是大量的、真实的，真实的数据可能是不完全的或者含有噪声的数据；其次，数据挖掘获得的信息或知识对于用户是有价值的；最后，发现的知识是能够被理解、被接受、被运用的，可以支持决策或能够支持特定的发现问题。

### 2.2 物联网数据挖掘存在的挑战

根据物联网数据特点，总结了物联网数据的特性对于数据挖掘技术提出的新的挑战，主要有以下几点：

（1）大量的物联网数据存储在不同的地点，因此通过中央模式很难挖掘分布式数据。

（2）物联网数据规模庞大，有大量的传感器节点，且需要实时处理，一般会采用中央结构，这样从很大程度上增加了对于中央节点的硬件要求。

（3）由于节点的资源是有限的，将数据放在中心节点的策略没有优化昂贵资源的使用，大多数情况下，中心节点不需要所有的数据，但是需要预估一些参数，可以在分布式节点中对原始数据进行预处理，再将必要信息传送给接收者。

（4）由于物联网数据存在许多外在因素，例如数据安全性、数据隐私、法律约束等。将所有数据统一存放在相同的数据仓库中的方式通常是不可行的。

由以上几点可以看出，对物联网进行数据挖掘时，现有的技术和方式存在很多弊端，需要进一步进行更深入的研究提出更多更好的解决方案。

## 3 基于云计算的物联网数据挖掘

该平台数据挖掘选用物联网数据集为例，选择目前研究热点 Hadoop 为基础平台搭建。平台主要包括四个大的模块：物联网感知层、传输层、数据层、数据挖掘服务层。具体如下：

（1）物联网感知层

感知层的作用主要是通过在目标区域内布置大量的采集节点，这些节点通过传感器、摄像头或其他仪器仪表来采集物联网数据，其中这些数据在物联网感知层内会存在通信，即存在无线传感器网络，通过这些网络汇聚数据到汇聚节点，然后对对数据进行汇总存储并且通过传输层最终传输到云平台数据中心。

（2）传输层

传输层主要是集传感器网络、无线网络、有线网络等多种

网络形态于一体的高速、无缝、可靠的数据传输网络，能够灵活快速的将感知数据传输至云计算数据中心，实现更加全面的互通互联；将各类监测设备进行联网数据传输，实现物联网中监测设备的网络化高速数据传输。

（3）数据层

数据层对于整个物联网数据挖掘平台是至关重要的，由于我们已经提到了物联网数据的异构性、海量性等特点，因此在数据层如和解决物联网这些数据存储及处理决定了物联网数据挖掘平台的可行性和性能。数据层主要包括两个重要模块：数据源转换模块、分布式存储模块。数据源转换模块主要用于物联网中异构数据的转换，分布式存储模块主要结合了 Hadoop 平台的文件系统 HDFS，采用分布式方式存储物联网海量数据。

由于在物联网中，不同的对象会有不同的数据类型来表示，甚至相同的对象都会用不同的数据来表示，因此数据源转换器的作用主要是来解决物联网数据异构性，它不仅可以保证数据存储的完整性，还能保证数据挖掘的顺利进行。数据源转换模块相当于数据层与感知层中各个监测设备的接口，并完成数据包解码以及按相应数据模型最终使分布式存储模块存储的都是有效并且完整的数据。数据转换器将不同类型的数据转换成 PML 数据，所以分布式存储在各个 NameNode 节点的文件类型为 PML[6] 类型数据。

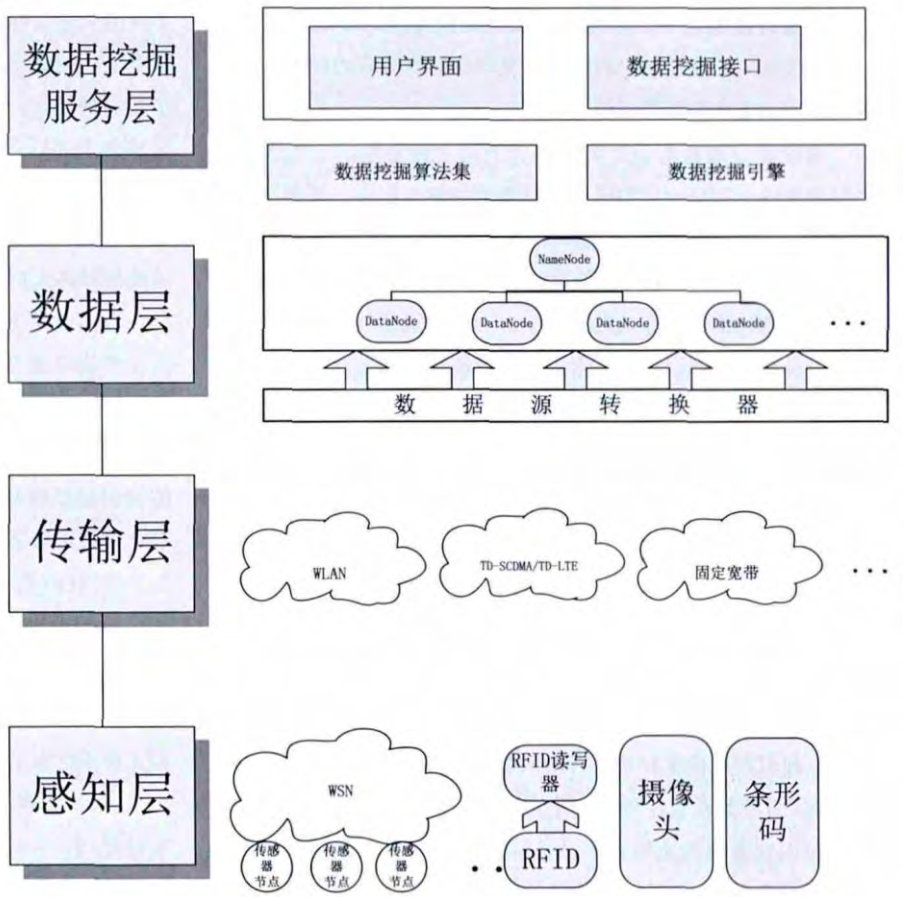


图 1 基于云计算物联网数据挖掘架构

Fig.1 The structure of data mining of IOT based on cloud computing

在此，我们提出 PML 的概念，PML 的出现提供一种通用的方式来描述自然物体，它是基于 XML 创建的语言，也有相同的核心思想。 PML 研发的目的是提供关于物品的详细信息，并促进物品信息的交换。例如，物联网的节点采集到信息，经过传输，在存储时利用 PML 进行建模，建模信息包括物体的属性信息、位置信息、单个物体所处的环境信息和多个物体所处的环境信息等，并包含了物体信息的历史元素，上述信息汇总后可以较为准确的描述物品的信息。

（4）数据挖掘服务层

数据挖掘服务层主要包括数据准备模块、数据挖掘引擎模块以及用户模块。数据准备模块主要包含了对于数据的清理、变换、数据规约等；数据挖掘引擎模块主要包含数据挖掘算法集、模式评估等; 用户模块主要包含数据挖掘知识的可视化表示。根据知识挖掘的类型不同，在数据挖掘引擎模块可以包括的功能主要有特征、区分、关联、聚类、局外者、趋势和演化分析、偏差分析、类似性分析等分类。提供这些功能的关键在于数据挖掘引擎模块中算法集提供各种功能的算法，而在 Hadoop 平台中数据挖掘算法需要对传统经典数据挖掘算法进行改进，即进行算法并行化处理。

用户模块是整个物联网数据挖掘平台直接面向使用人员的部分，所以应该具有良好的友好性，用户可以通过界面操作进行数据挖掘任务，并能够得到可以被理解的知识。为了增强平台的可移植性，在用户服务底层模块增加开放接口，从而可以使第三方调用物联网数据挖掘平台的功能，使物联网应用更加丰富。( 图 1)

## 4 实验验证

### 4.1 基于云计算物联网数据挖掘平台工作流程

数据挖掘流程如下图所示，用户请求进行数据挖掘，主控节点接收到请求后会判断是否可以进行该任务，并返回给用户发送该任务是否可以进行，如果可以进行，主控节点在数据挖掘算法集存储模块中调用用户所需的数据挖掘算法，此时，HDFS 文件存储系统中的数据文件会进行数据规约等处理。此后根据数据挖掘算法进行分布式数据挖掘，分布式数据挖掘的思想是 MapReduce，它采用了 Master/Slave( 主 / 从 ) 结构。通过主节点将数据挖掘任务进行划分后，传递到需要完成具体工作的从节点上，这些节点负责具体去处理数据挖掘的具体数据。JobTracker

负责 Job 和 Tasks 的调度，而 TaskTracker 负责执行 Tasks。

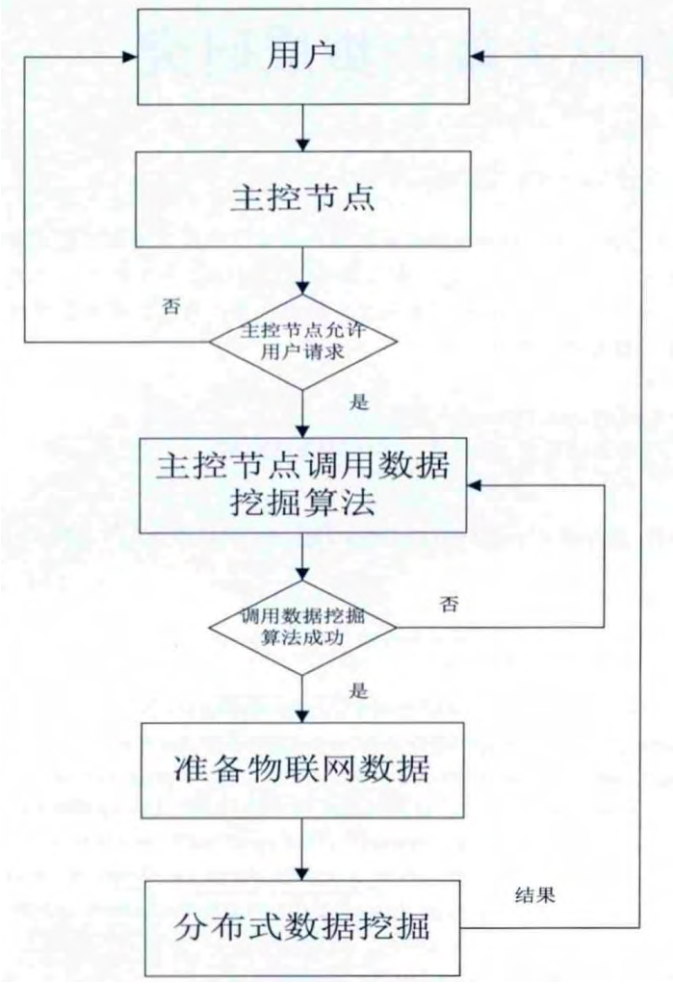


图 2 物联网数据挖掘流程图

Fig.2 The workflow of data mining in IOT

## 4.2 实验验证

本文通过搭建 Hadoop 平台[7]，并将通过数据转换器转换成 PML 格式的数据进行分布式存储后，运行经过 MapReduce 化的数据挖掘算法（此实验选用了 Apriori 算法）[8]，分析整个平台的可行性和性能。

4.2.1 实验环境

本实验选用了一台 PC 机（配置为 2G 内存，250G 硬盘，系统为 Windows7）上安装了安装虚拟机的方式部署多个分布式节点，一共安装了 3 个虚拟机，操作系统均为 Linux 系统[9]（1 个 NameNode，2 个 DataNode）。

另外还安装了 Linux 版本的 Eclipse 7.5 集成开发环境，并在本机 Windows7 下安装了 SSH Secure Shell Client 方便实验时传递数据使用，每个虚拟机的操作系统下也安装了 SSH 服务，并且会进行一些基本的设置，在运行 Hadoop 时需要用到的。

4.2.2 实验过程

配置完成 Hadoop 平台后，选取了一组用于关联规则算法的实验数据，将实验数据通过 C++ 代码编写的程序通过关键字搜索方式转换成标准类型的 PML 文件（大小为 1 G），将文件通过 HDFS 的命令放入到 Hadoop 平台上进行分布式存储。运行经过改进后的 Apriori 算法（Java 语言编写），得到运行结果，查看是否找到了实验数据集中的所有频繁项集。

另外，我们还会选取不同大小的文件进行上述实验，并对比运行时间等，用于分析平台的性能。

4.2.3 实验结果

在验证了该平台的可行性后，通过运行不同大小的数据集，得到的运行时间如下表所示：

表 1 文件大小与运行时间对应关系

Tab.1 File Size and Run-time

| 文件大小（M） | 运行时间（s） |
|---|---|
| 200 | 109 |
| 500 | 210 |
| 750 | 293 |
| 1000 | 312 |

从上表可以看出随着数据量增大，改进后的 Apriori 算法呈现线性增加，可以在数据量变大的情况下完成频繁项集的发现，因此，可以看出该平台有很好的扩展性，能够满足物联网海量数据的挖掘。

## 5 结论

本文对于提出了物联网数据挖掘对于物联网产业的发展有十分重要的意义，由于物联网的特点决定物联网数据挖掘存在许多困难，为了解决这些困难，对于物联网数据挖掘和云计算结合进行了许多研究，并提出了结合 Hadoop 平台进行分布式数据挖掘的观点，通过实验验证了这种思路的可行性。

## 参考文献

[1] 张福生 . 物联网 [M]. 山西：山西人民出版社，2010.

[2] 张为民，赵丽君 . 物联网与云计算 [M]. 北京：电子工业出版社，2012.

[3] S. Haller, S. Karnouskos, and C. Schroth.The Internet of Things in an enterprise context[J].Future Internet Systems (FIS), LCNS, vol. 5468. Springer,2008,14-8.

[4] Apache. Welcome to apache hadoop[OL].2010.http://hadoop.apache.org/.

[5] 汤为 , 孙才红 . 数据挖掘在 FAST 动态监测系统中的应用 [J]. 软件 ,2013,34(6):31-34

[6] 黄玉兰著 . 物联网核心技术 [M]. 北京：机械工业出版社，2011.

[7] 拉姆 .Hadoop 实战 [M]. 韩冀中 . 北京：人民邮电出版社，2011.

[8] 贺海梁，袁玉宇 . 基于 REST 的面向资源 Web 应用架构参考模型 [J]. 软件，2012,33（11）:59-63

[9] 夏兆彦 .Linux 指令范例速查手册 [M]. 北京: 清华大学出版社 ,2011.

[10] 张沙清 , 郭建华 , 杨玉法 , 等 . 基于物联网的猪肉产品质量安全监管与溯源系统 [J]. 软件，2013，34（12）：6-9.