植物遗传资源学报 2015,16(3):649-652
Journal of Plant Genetic Resources
DOI:10.13430/j.cnki.jpgr.2015.03.031

基于云计算的农作物种质资源数据挖掘平台研究

潘　恺,方　汎,陈丽娜,曹永生
(中国农业科学院作物科学研究所, 北京 100081)

摘要:针对作物种质数据量大、多维带来的挖掘效率偏低的问题,通过探讨云计算技术及其解决方案,提出了一种基于Hadoop的农作物种质资源数据挖掘平台,详述了平台的各功能模块并给出了具体的开发方案。通过改进经典的Apriori算法并在平台上对其进行效率测试,验证了平台的有效性、可行性。

关键词:种质资源;云计算;Hadoop;Apriori算法

Research of Crop Germplasm Resources Data Mining Platform Based on Cloud Computing

PAN Kai, FANG Wei, CHEN Li-na, CAO Yong-sheng
(*Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081*)

Abstract: The massive volume and multiple dimensions of germplasm data have caused the low efficiency of data mining. This problem is envisaged in this paper by presenting a Hadoop-based data mining platform for crops germplasm after detailed analysis of cloud computing technology and its solution. Different functional modules will also be discussed in detail, based on which, specific development programs of the platform are also to be proposed. The efficiency and feasibility of the platform will be verified through efficiency tests of improved classic Apriori algorithm.

Key words: germplasm resources; cloud computing; Hadoop; Apriori algorithm

目前,我国已建成了拥有200种作物、42万份种质信息、2400万个数据项值的中国作物种质资源信息系统(CGRIS, Chinese Crop Germplasm Resources Information System)[1-2],随着考察、收集、评价、鉴定工作的推进,种质资源数据呈现不断增长的趋势,如何帮助种质资源工作者和育种家更好地利用这些数据逐渐成为种质资源信息领域研究的焦点。

数据挖掘[3-4]作为数据处理和知识发现的一种重要技术,能够从海量数据中挖掘出具有决策价值的知识。在种质资源领域,数据挖掘也有一定应用,如杉木伴生树种发现[5]、杨树无性系叶片"原子簇"分析[6]等。然而,由于种质数据具有量大、多维的特点,传统的挖掘算法,如Apriori、K-means等,在应用时普遍存在效率偏低的问题。

大数据时代下,云计算[7-9]技术的发展为海量数据分析提供了廉价的解决方案,也为挖掘平台提供了新的发展方向。一般认为,云计算是一种具有高可扩展性、使用虚拟资源、可供用户共享特点的计算模式。目前的云计算模型有很多,但出于商业考虑,这些模型大都属于细节保密,而Hadoop[10-12]作为开源的云计算模型,其主要优点有扩容能力强、成本低廉、效率高、高可靠性、免费开源及良好的可移植性。基于此,本研究提出了一种基于Hadoop云计算技术的农作物种质资源数据挖掘平台,并通过将经典的Apriori关联规则算法移植到平台上,验证了平台的有效性、可行性。

1　数据挖掘平台设计

1.1　平台设计目标与原则

平台的总体目标是发挥Hadoop对海量数据的

收稿日期:2014-10-23　　**修回日期**:2014-12-01　　**网络出版日期**:2015-04-10
URL:http://www.cnki.net/kcms/detail/11.4996.S.20150410.1613.007.html
基金项目:国家农作物种质资源平台(2005DKA21001)
第一作者研究方向为种质资源信息系统、数据挖掘。E-mail:pankai@cgris.org
通信作者:曹永生,主要从事种质资源信息系统、图像识别、GIS领域研究。E-mail:caoyongsheng@caas.cn

处理优势,满足大数据时代下种质数据不断增长的计算能力需要,方便用户利用数据挖掘相关方法获取海量种质数据中潜在的知识、信息。

平台设计的原则包括:经济原则,利用多台服务器搭建云平台以整合其闲置的计算能力;高效原则,查询效率与挖掘效率兼顾,并将挖掘过程中计算量大的模块扩展到 Hadoop 集群的各个节点上,以提高计算效率;专业原则,平台面向种质资源领域相关研究人员,各模块需充分考虑种质数据的特点,提高针对性。

1.2 平台框架设计

根据平台的设计目标与原则,本研究提出的种质资源数据挖掘平台框架如图 1 所示,该框架主要包含以下模块。

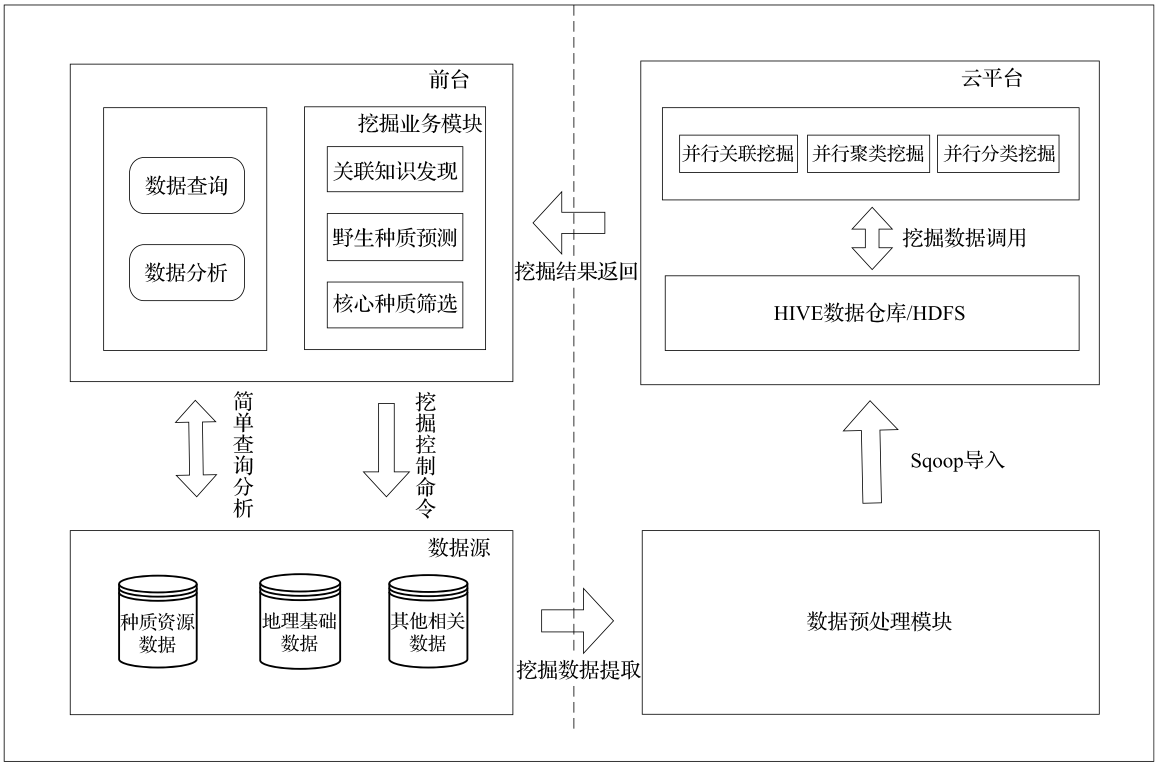

图 1 平台框架图

Fig. 1 Platform frame

1.2.1 数据源 包含种质资源数据、地理基础数据和其他相关数据 3 大类数据,数据类型除传统结构化数据外,还包含栅格数据、矢量数据等非结构化数据。由于 Hadoop 在处理大量数据的运算时才具备优势,因此对于简单的查询分析请求,由服务器直接进行响应,并将结果以可视化方式在前台展示。对于复杂的数据挖掘请求,服务器会从数据源取出相关挖掘数据,交由数据预处理模块进一步处理。

1.2.2 数据预处理模块 负责将原始数据转换成适用于挖掘算法的形式。不同的挖掘算法对于数据的形式有不同的要求,以关联规则挖掘算法为例,算法要求待挖掘数据为离散的事务型数据,而种质数据以表格形式存放在数据库中,这就需要先进行表格型数据到事务型数据的一个转换,如果待挖掘的种质属性为连续型数据,还需进行连续数据离散化的操作。

1.2.3 云平台 为挖掘平台的核心模块,提供对海量数据的分布式存储与计算能力。云平台由若干台服务器组成,采用 Hadoop 云计算模型,其架构如图 2 所示,其中一台服务器作为主节点,负责海量数据的存储管理和计算控制,剩余服务器作为子节点,负责切片后数据的存储和计算执行。挖掘算法根据 Map/Reduce 框架编写,重点是实现算法的并行化,将算法里计算量大的部分部署在各子节点上进行计算,从而使平台处理海量数据的能力大大增强。

1.2.4 挖掘业务模块 在平台挖掘能力的支撑下,为种质研究人员提供面向海量数据的挖掘业务,包括关联知识发现、野生种质预测、核心种质筛选等,分别应用了挖掘算法中的关联、分类和聚类算法。

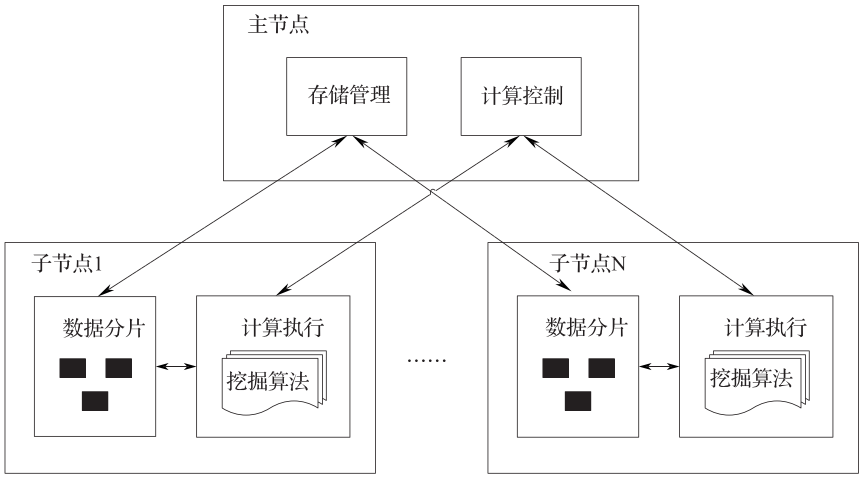

图 2　云平台架构

Fig. 2　Cloud platform architecture

其中关联知识发现不仅可以挖掘种质特性间的关联规则,还有能力挖掘种质特性与气象、地理等海量外部数据间的关联规则。野生种质预测则是依据平台收集的野生种质数据与气象、地理数据建立训练集,来预测未知区域野生种质所具备的特性。而利用平台已收集的大量相关数据也可以更全面、精确地筛选出核心种质。

1.3　开发方案

挖掘平台采用典型的 B/S 架构,开发环境基于 Linux 系统。平台的各功能模块由 J2EE 开发工具和脚本语言共同实现,确保平台兼具稳固性和灵活性。挖掘算法应用 JAVA 语言编写,从而使挖掘算法与 Hadoop 具有更好的一致性。

2　挖掘算法改进

目前,数据挖掘领域主要有关联、聚类和分类 3 大类算法,其中关联规则算法侧重于发现数据各属性间的相互关系,是数据挖掘中最成熟、最重要的研究内容之一。本研究以经典的 Apriori[13-14] 关联规则算法为例,根据 MapReduce 计算框架,将其移植到 Hadoop 平台上并对其有效性进行测试。

2.1　传统 Apriori 算法

Aprioir 算法又称先验算法,它采用逐层搜索的方法来产生频繁项集,即利用频繁 k 项集 L_k 产生 L_{k+1}。算法首先扫描数据集,找出满足最小支持度集合,记为 L_1,L_1 与自身连接、剪枝后产生 2 项候选集 C_2,扫描数据集,找出 C_2 中满足最小支持度的集合,记为 L_2,相同步骤,由 L_2 产生 L_3,如此循环迭代,直到不产生新的频繁集为止。其中,每次循环都会扫描一次数据集,这将产生大量的系统开销,是限制算法对海量种质数据进行挖掘的关键步骤。

2.2　基于 MapReduce 的 Apriori 算法改进

针对 Apriori 算法存在的问题,采用 MapReduce 框架“分割、合并”的思想对其进行改进,首先对海量数据集均匀分片,将分片后的 N 个数据片段分发到集群的 N 个子节点上,接着依次执行 Job_1,Job_2……Job_k 来生成相应的频繁集 L_1,L_2……L_k,具体算法过程如下。

(1)Job_1 采用类似单词计数的过程,并行扫描数据集,统计满足最小支持度的 L_1,将 L_1 写入 HDFS 中。

(2)Job_2 从 HDFS 中读取 L_1,L_1 与自身连接、剪枝后生成 C_2,并行扫描数据集,统计 C_2 各集项的个数,筛选出满足最小支持度 L_2,将 L_2 写入 HDFS。相同步骤,由 Job_k 生成 L_k,直到没有新的频繁集生成为止。

改进后的 Apriori 算法实现了在 Hadoop 集群上的并行化计算,避免了对海量数据集的重复扫描,而只需各子节点扫描分片后的数据片段,这极大地提高了算法的效率。同时,对海量数据集的均匀分片也解决了 Hadoop 容易产生的各子节点负载不平衡的问题。

3　平台性能测试

3.1　试验环境

本研究利用 VMware 10.0 虚拟机建立计算机集群,用于测试的 Hadoop 集群由 4 台配置相同的计算机组成,通过 SSH 协议进行互连,其中 1 台作为主节点(Master)负责数据管理与任务调度,3 台作为子节

点(Slave)负责数据存储和任务执行。每台计算机配置为:处理器类型为 Intel(R)Xeon(R)X5550@2.67 GHz,内存容量为 2GB。操作系统使用 Ubuntu12.04,Hadoop 使用 1.2.1 版。

3.2 测试内容与结果

试验选取水稻种质数据表中的 10 项特性属性作为测试数据,包括籼粳、早中晚、水陆、粘糯、米色、芒长、粒形状、粒长度、颖尖色和颖壳色,在 Hadoop 计算平台上应用改进后的 Apriori 算法,挖掘支持度大于 0.6,置信度大于 0.6 的关联规则,挖掘出的部分关联规则如表 1 所示。

表 1 关联规则结果

Table 1 The results of association rules

序号 No.	支持度(%) Support	置信度(%) Confidence	关联规则 Asocciation rules
1	64.4	94.0	[粒形状-椭圆]⇒[水陆-水]
2	75.5	93.5	[粘糯-粘]⇒[水陆-水]
3	71.3	93.4	[米色-白]⇒[水陆-水]
4	60.0	93.4	[籼粳-籼,水陆-水]⇒[粘糯-粘]
5	61.1	93.1	[籼粳-籼]⇒[粘糯-粘]
6	71.6	93.0	[芒长-无]⇒[水陆-水]

通过对几组不同规模的数据进行挖掘,比较 Apriori 算法在单机模式和 Hadoop 集群上的运算时间,两种模式下挖掘消耗的时间如表 2 所示。

表 2 试验结果统计分析

Table 2 The experimental results of statistical analysis

试验序号 Test No.	数据量(Mb) Data size	单机耗时(s) Standalone time-consuming	平台耗时(s) Platform time-consuming
1	8	5	5
2	15	9	8
3	30	16	15
4	60	28	25
5	120	内存溢出	56

上述试验结果表明,当数据量较小时,并行化的 Apriori 算法并没有明显的优势,这是因为 Hadoop 集群上 Job 的启动与交互需要消耗一定的资源,但当数据量较大时,单机模式下的内存开销增加至溢出,导致无法完成挖掘任务,而集群上 Apriori 分布式计算的优势便逐渐体现出来,可以完成挖掘任务,证明本研究提出的挖掘平台对于处理海量多维种质数据是有效的。

4 结语

大数据时代下云计算的快速发展为海量种质数据的挖掘提供了新的研究方向,而 Hadoop 开源、细节透明的特性为挖掘云平台的搭建提供了切实可行的技术支持。试验表明,本文提出的基于 Hadoop 云计算技术的农作物种质资源数据挖掘平台是有效可行的,在下一步研究中,需进一步提高平台挖掘算法的效率,丰富平台挖掘业务,扩大平台在种质资源领域的应用范围。

参考文献

[1] 曹永生,方沩. 国家农作物种质资源平台的建立和应用[J]. 生物多样性,2010,18(5):455-456

[2] 方沩,曹永生. 中国作物种质资源信息系统[J]. 科研信息化技术与应用,2012,3(6):66-73

[3] Chye K H,Chin T W,Peng G C. Credit scoring using data mining techniques[J]. Singapore Management Review,2004,26(2):26-47

[4] 王光宏,蒋平. 数据挖掘综述[J]. 同济大学学报,2004,32(2):246-252

[5] 黄世国,林思祖,林大辉. Apriori 算法在杉木伴生树种选择中的应用[J]. 福建农林大学学报:自然科学版,2008,37(1):70-72

[6] 段旭良. 杨属种质资源数据挖掘研究[D]. 北京:北京林业大学,2008

[7] 陈全,邓倩妮. 云计算及其关键技术[J]. 计算机应用,2009,29(9):2562-2567

[8] 何清. 大数据与云计算[J]. 科技促进发展,2014,10(1):35-40

[9] Ercan T. Effective use of cloud computing in educational institutions[J]. Procedia-Social and Behavioral Sciences,2010,2(2):938-942

[10] Apache. Welcome to Apache Hadoop[EB/OL]. [2014-09-27]. http://hadoop.apache.org/

[11] 王宏宇. Hadoop 平台在云计算中的应用[J]. 软件,2011,32(4):36-38

[12] 金松昌,方滨兴,杨树强,等. 基于 Hadoop 的网络安全日志分析系统的设计与实现[C]//全国计算机安全学术交流会论文集. 北京:中国科学技术大学出版社,2010:257-262

[13] Agrawal R,Srikant R. Fast Algorithms for Mining Association Rules[C]//Proceedings of the 20th International Conference on Very Large Databanses,Santaigo,Chile,1994:487-499

[14] 黄立勤,柳燕煌. 基于 MapReduce 并行的 Apriori 算法改进研究[J]. 福州大学学报:自然科学版,2011,39(5):680-685