文章编号：2095-2163(2021)01-0167-03 中图分类号：TP311 文献标志码：A

面向云计算环境下 Web 数据挖掘技术

曾展挺
（惠州城市职业学院，广东 惠州 516025）

摘 要：在云计算环境下，Web 数据挖掘技术得到了快速发展。由于云计算的应用，Web 数据挖掘体系已体现出新的特点。分析云计算环境下 Web 数据挖掘技术的特点，可以明确应用要点，可以实现云计算在数据存储中的突破，实现存储的能力与安全性的提高。从海量数据中高效挖掘有价值的资源，属于信息技术要解决的关键问题。云计算技术支持下的数据挖掘实现了资源的优化配置，体现出实用性、虚拟性的特点，可以保证数据挖掘的高效、精准。因此，有必要构建基于云计算的数据挖掘模式，保证数据挖掘具有更高的精准度，并实现挖掘成本的降低。

关键词：云计算；技术应用；数据挖掘；Web 数据

Web data mining technology in cloud computing environment

ZENG Zhanting
（City College of Huizhou, Huizhou Guangdong 516025, China）

【Abstract】In the cloud computing environment, Web data mining technology has been developed. Due to the application of cloud computing, Web data mining system has new characteristics. By analyzing the characteristics of Web data mining technology in the cloud computing environment, the application points can be clarified, the breakthrough of cloud computing in data storage can be realized, and the storage capacity and security can be improved. Mining valuable resources efficiently from massive data is the key problem to be solved in information technology. Data mining supported by cloud computing technology realizes the optimal allocation of resources, reflects the characteristics of practicality and virtuality, which could guarantee the efficiency and accuracy of data mining. Therefore, it is necessary to build a data mining model based on cloud computing for ensuring higher accuracy of data mining and reducing mining cost.

【Key words】cloud computing; technical application; data mining; Web data

当前互联网技术发展迅猛，互联网信息也呈现持续高速增长态势，如何由海量数据中发现有价值的信息即已成为数据挖掘技术的研究热点。研究可知，Web 数据挖掘是对 Web 海量数据加以分析，借助数据挖掘算法筛查出有价值的信息，而这些信息对于诸如趋势走向预测和商业行为决策等是十分有用的。对此拟展开如下研究论述。

1 基于云计算的 Web 数据挖掘体系

在互联网中，运用数据挖掘可以将 Web 划分为不同的节点，借助云计算技术实现 Web 中不同节点的关联，建立起数据挖掘体系。在应用实践中，主控节点要实现客户端与不同节点的网络连接；算法节点可以为数据挖掘的应用提供算法支持，对其可理解为算法仓库；数据节点作为数据存储的数据库；服务节点是执行系统下达的指令，并对计算结果加以反馈。针对 Web 数据挖掘的设计实现，本次研究中是将体系分为 4 个层面，详见图 1。该体系中，每个层面的定制功能可做阐释分述如下。

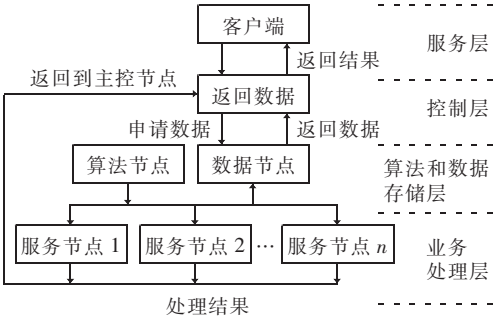

图 1 云计算环境下 Web 数据结构

Fig. 1 Web data structure in the cloud computing environment

（1）服务层。借助 Web 数据挖掘来提取数据，将信息传送给用户。

（2）控制层。通过主控节点对返回的结果进行分析，同时确定算法的有效性，用来实现更深层次的数据挖掘。

（3）算法和数据存储层。存储反馈的数据，包括初始数据、挖掘后数据，避免数据、算法发生丢失，造成损失。一旦发生意外，系统还可从数据存储区中重新找回数据进行恢复。

作者简介：曾展挺（1980-），男，硕士，讲师，主要研究方向：人工智能应用。

收稿日期：2020-10-30

(4)业务处理层。借助程序重新对存储层数据加以分配,并借助服务节点将信息反馈到主要控制节点。

2 云计算的 Web 数据挖掘

2.1 云计算的数据挖掘概述

在网络技术快速发展,以及在多领域应用普及的背景下,就产生了海量的数据,Web 数据挖掘技术也随即获得了广泛的应用。当前 Web 数据技术已趋于成熟,并与云计算技术的应用密切相关。借助云计算技术,数据信息的高效处理、分析已然成为可能,数据挖掘的服务性和时效性也变得更好,数据的利用价值也更加突显。数据挖掘过程中,先要对数据加以处理,再借助数据分析,通过算法得到数据的评价和表达,成功提取到有价值的信息。

云计算环境下 Web 数据挖掘技术采用分布并行处理方式,具体特点可做分析阐述如下。

(1)云计算环境下的 Web 数据挖掘可以保证更高的效率,过程中应用了并行处理方式,提升了海量数据的挖掘速度。在云计算环境下,可以为不同要求的客户提供个性化服务,且服务成本也会更低,有利于数据挖掘的快速实现。对于中小客户,可以无需考虑使用大型高端服务器。

(2)云计算环境下的数据挖掘通过块划分自动分配计算任务,保证节点加载的灵活性。

(3)云计算环境下的数据挖掘技术有着良好的用户体验。普通用户只需登录云服务平台即可,而特殊的用户则可以通过个性化的数据服务来满足其实际需求。

(4)云计算环境下的数据挖掘可进行动态增删,还可自由添加结点,这就提升了海量数据的处理速度,设备的利用率也得到了同步提升。

2.2 云计算环境下的数据挖掘实现方式

2.2.1 建立数据挖掘模型

数据挖掘模型的建立要结合客户的实际需求。针对商业客户而言,则需要借助技术优化模式从海量 Web 数据中挖掘出具有商业价值的实用数据。因此数据挖掘模型建立时要确保真实性、合理性。尤需指出的是,云计算技术的大范围应用,实现了大容量存储,提升了并行处理能力,有效解决了常规模式下数据挖掘存在的制约问题。

建立数据挖掘模型,还要结合 Web 挖掘建立流程。数据挖掘存在多种影响因素,这种方式有别于传统的挖掘模式。Web 数据挖掘流程的设计要考虑到诸多不利因素:Web 数据挖掘技术融合 Web 网页并不是各类技术的简单叠加,而是包含对信息检索、选择并初步处理信息、找到模式且加以分析等在内的一系列步骤。对于 Web 信息的检索,就是通过爬取网站新闻、日志等数据信息,对其加以甄别和筛选,滤除无价值的信息,并初步处理有价值的数据。再对处理后数据进行筛选和验证,完成有价值的数据提取。研究发现在云计算技术的支持下,数据挖掘时可以采用流程化的模式:向模块依据用户需求发出指令,指令上传到云服务器,服务器会自动识别出该指令,调取已存储的数据,引用最优算法,对数据进行预处理,在此基础上反馈到云平台,运行结果则将采用可视化的方式来呈现给用户。云计算环境下,若要提升基础架构库的可靠性,则亟需建立安全可靠的流程,从而保证最终的服务效果。服务流程的设计要有利于规范数据挖掘,流程要结合差异化的用户需求,同时结合数据挖掘的目标,体现出技术基础架构库的优势,降低对人的依赖性。

2.2.2 算法的设计方式

在云计算环境下,挖掘数据可运用 SPRINT 算法,依据设置流程,先创建决策树,然后剪枝。创建决策树时,要对数据加以筛分,剪枝时则是去除无用的数据。SPRINT 算法设计时也融入了不同数据的特征。划分属性表后,节点分裂了,可以确定属性表。属性表包含了索引、类,放置于内存空间外,表明了节点属性。对于数据处理,不间断的刷新即可获得最有效的分裂点。如果采用离散型,可以借助直方图来表达属性值的分布特点。算法设计中的并行处理则可保证算法的运行效率。引入哈希表,存储中不同节点发生分裂后子节点的数据变化也可以直观呈现出来,即使得对节点实施的并行处理就具备了分割依据。应用哈希表体现出决策节点号码的信息以及树节点子信息。算法移植后,通过 MapReduce 算法可以进行优化,算法的应用可以快速创建出决策树,这样就提升了算法执行效率。

2.2.3 数据挖掘算法的应用

在 Web 数据挖掘技术中,至关重要的数据结构是 Web-Graph。该技术可以描述 Web 信息,并可广泛应用于社交网络、搜索结果排序、网络爬虫等场合。Web-Graph 对 Web 链接进行分析是基于图论算法的应用,因此为数据处理分析提供了有利条件。对于算法数据结构的数据,采用 Web-Graph 算法分析数据时,先要明确数据描述算法,通常可以采用矩阵法来描述数据,还要结合行列特点,排列节点数

据，从而形成网络矩阵。网络中的矩阵阶数可用节点数进行表示。算法体现了网页的链接关系，其关系则借助矩阵来进行描述。对于矩阵的创建，数据表达出行、列节点之间的联系。对于取值，数据的矩阵元素可表达出一定的差异，可以表达各个节点 Graph。利用 Graph 的差异，体现出社交平台中的用户关系。在社交网络中，用户信息交换存在双向关系。用户在得到相互认可后，才会确认为好友，因此利用数据结构，就可采用对称矩阵的形式来描述用户的关系。在数据应用中，借助二维数组来表达矩阵，如果应用高级语言去处理 Graph，矩阵采用的就是数据结构。

GraphML 应用存储具有可靠性、长期性的特点。GraphML 作为通用文件格式借助 XML 语言对图形特征加以描述。目前，许多开发语言都能够解析 GraphML，因此 Graph 的生成、处理、存储等在很多场景中都可以成功得到运用。GraphML 还表现出简单、直观等优势，为开发人员提供了多方面的便利。不仅降低了数据挖掘难度，有利于开发人员的后续修改，而且为程序应用创建了良好的数据条件。Graph 数据常用结构包括分级图、超图、无向图等。在数据挖掘过程中，就是通过爬取得到页面信息来详细分析 Web 连接，从而形成 Graph 结构。此类挖掘算法的优势就是易于实现。而在分析文本的页面链接时，会消耗计算资源，除 Web 关联外，利用 Graph 结构，还可以描述常见事物的关联。

针对数据收集，传统的方式是直接收集互联网上的数据，存储于数据仓库中。但是数据仓库中的数据却可能发生丢失。在云计算技术下，数据收集时会首先筛选互联网上的海量信息数据，经数据转化生成半结构化的文件，再将其保存于分布式系统中。针对数据处理，是由云计算中的任务主节点来实现整体的统筹控制。任务主节点会对任务进行分类细化，并将其有针对性地分配到互联网上的空闲计算机加以处理。接下来再将网络中分散中心处理后的信息在集结汇总后，一并传送到主节点。这种方式高效地利用了计算机资源，并且保证了数据处理效率。

3　结束语

面对海量的网络信息，Web 数据挖掘体现出极高应用价值。云计算的应用为 Web 数据挖掘的实现创造了有利条件。云计算下的 Web 数据挖掘保证了网络资源的实时分析与处理，数据挖掘的效率也得到提升。

参考文献

[1] 朱娜. 基于云计算技术的数据挖掘平台设计与实现[J]. 信息记录材料，2018，19(6)：79-81.

[2] 葛晓玢，刘杰. 基于云计算的数据挖掘平台架构及其关键技术研究[J]. 景德镇学院学报，2017，32(3)：26-29.

[3] 薛医贵. 云计算在 WEB 数据挖掘技术中的应用研究[J]. 自动化与仪器仪表，2017(5)：156-157，161.

[4] 熊伯安. 基于大数据时代的数据挖掘及分析[J]. 电子世界，2016(20)：121，123.

(上接第 166 页)

析，得到了各因素单独及交互作用对翘曲的影响。

(3)依据响应面法建立的模型，构建了遗传算法的适应度函数，建立了遗传算法模型，在 Matlab GUI 中迭代优化得到最小翘曲变形量为 0.030 618 mm 及此时参数组合，即：模具温度 160 ℃，熔体温度80 ℃，注塑时间 2 s，保压压力 84.9 Mpa。

(4)根据最小翘曲变形量，对模具反变形补偿，按照优化后获得的最佳工艺参数进行封装验证，芯片翘曲满足评价标准，符合生产要求。

参考文献

[1] SUHIR E. Predicted bow of plastic packages of integrated circuit (IC) devices [J]. Journal of Reinforced Plastics and Composites, 1993,12:951-972.

[2] MIREMADI J. Impact of PBGA-ball-coplanarity on formation of solder joints [C]// 45th Electronic Components and Technology Conference. Las Vegas, NV, USA: IEEE,1995:1039-1050.

[3] HU K X, YEH C P, DOOT B, et al. Die cracking in flip-chip-on - board assembly [C]// 1995 Proceedings Electronic Components and Technology Conference. Las Vegas, NV, USA: IEEE, 1995:293-299.

[4] 陈艳霞，陈如香，吴胜金. Moldflow 2012 完全自学手册[M]. 北京：电子工业出版社，2012.

[5] 吴其晔，巫静安. 高分子材料流变学[M]. 2 版. 北京：高等教育出版社，2014.

[6] 曹阳根，傅意蓉，王元彪，等. IC 封装模流道平衡 CAE 应用[J]. 模具工业，2004(4)：42-44.

[7] 刘丽平. 空调面板注塑模充模流动分析及工艺参数优化[D]. 太原：太原理工大学，2010.

[8] 张帅，陈振，尹泽康，等. 基架注塑成型数值模拟及工艺优化[J]. 塑料工业，2017，45(2)：65-70.

[9] 齐雪，廖秋慧，祝璐琨，等. 基于响应面法的汽车接插件注塑工艺优化[J]. 塑料科技，2018，46(10)：95-99.

[10] SHEN Y K, YE T W , CHEN S L , et al. Study on mold flow analysis of flip chip package[J]. International Communications in Heat and Mass Transfer, 2001, 28(7):943-952.

[11] 郑树泉，王倩，武智霞，等. 工业智能技术与应用[M]. 上海：上海科学技术出版社，2019.

[12] 段家现，闫西坡. 基于反变形和 CAE 的手机前壳翘曲优化[J]. 轻工机械，2016，34(6)：93-97.