基于改进型迭代算法的web数据关联规则挖掘

刘啸[1],刘玉龙[2]

1. 江苏师范大学现代教育技术中心,徐州 221116
2. 江苏师范大学计算机科学与技术学院,徐州 221116

摘要 因特网上的数据越来越多、越来越复杂,这些异构、动态、分布的信息使得传统数据挖掘方式已经不能达到实际要求。本文提出了一种面向web数据挖掘的改进型迭代算法,将迭代方法与多服务器并行算法进行结合,并采用该算法建立了一个支持并行关联规则的web数据挖掘模型,融合存储节点本地计算的思想。实验证明,该模型能够提高web数据挖掘的效率,并有随着数据量增加执行率升高的特点。

关键词 web挖掘;迭代算法;并行算法;本地计算

中图分类号 TP311 **文献标志码** A **doi** 10.3981/j.issn.1000-7857.2015.03.015

Web data mining of association rules based on an improved iterative algorithm

LIU Xiao[1], LIU Yulong[2]

1. Modern Education Technology Center, Jiangsu Normal University, Xuzhou 221116, China
2. School of Computer Science & Technology, Jiangsu Normal University, Xuzhou 221116, China

Abstract With the increasing dependency of all aspects of social life on Internet, the data on the internet is becoming more and more massive, and also more complex. This heterogeneous and dynamic information which is also distributed makes the traditional data mining unable to achieve actual requirements. This paper proposes an improved iterative algorithm for web data mining: combining iteration method with a parallel algorithm. And a web data mining mode is set up by the algorithm with the idea of local computing of storage nodes, which supports the parallel association rule. Experimental results show that this mode can improve the efficiency of web data mining and its implementation rate will rise as the data quantity increases.

Keywords web mining; iterative algorithm; parallel algorithmic; local computing

近几年,网络技术的发展使得社会生活的各个方面都离不开网络信息。web数据挖掘的主要功能即是实现网络数据的智能化处理,从而能够利用有效的数据挖掘技术,收集、获取感兴趣的信息,得到和抽象出大量信息的关系模型,挖掘出更深层次的信息。从近年来社会生活发展的各个方面来看,信息挖掘对个人、企业等都有着很重要的意义,例如通过网络挖掘,企业能够得到更多有效的网络信息,快速对各项信息进行收集和抽象,从而为研究决策提供有力的支持[1]。

数据挖掘是一项复杂的数据处理技术,而由于网络数据具有分布式、动态性、异构性等特点,使得对网络信息的挖掘

收稿日期:2014-08-13;修回日期:2014-10-13
基金项目:江苏省高校自然科学研究基金项目(11KJB450001);江苏师范大学自然科学基金项目(14XLB04)
作者简介:刘啸,讲师,研究方向为计算机网络、数据挖掘、信息安全,电子信箱:lllxiao@126.com;刘玉龙(通信作者),教授,研究方向为离散数学与算法分析、信息安全,电子信箱:ylliu@163.com
引用格式:刘啸, 刘玉龙. 基于改进型迭代算法的web数据关联规则挖掘研究[J]. 科技导报, 2015, 33(3): 90-94.

更加有难度[2]。如何将全球范围内庞大的数据进行有效集合,实现高效挖掘成为近年来研究的热点。复杂度极高的数据处理和高性能挖掘算法,是目前亟待解决的问题[3]。本文通过引入本地计算思想,将迭代式的数据挖掘算法进行扩展。使用该数据挖掘算法,研究和设计了一种基于此算法的数据挖掘模型,试图解决分布式网络系统的资源挖掘效率提升问题,得到更加优良的挖掘性能。

1　web挖掘

1.1　挖掘对象分析

从研究对象看,由于web数据更多的是以网络日志的形式出现,因此,本文主要考虑通过对网络日志的挖掘,发现潜在规律性的用户行为,而典型的MapReduce模型使用范围就在web数据挖掘的词频挖掘中,适合web日志关联规则挖掘。网络日志包含了绝大多数用户倾向性数据,例如浏览请求地址、用户信息特征、页面链接关联程度等,而这些数据对数据挖掘、信息判定等理论起着支撑作用[4]。通过网络日志对用户行为模式进行分析的处理方法逐渐受到所重视,在电子商务、网络新闻、传媒监测等方面开始得到应用。

根据不同需求开发人员通常使用不同的分析方法,完成对网络日志的信息采样。需要分析某一用户的特征行为时,例如购买特征、喜好倾向性等,往往先要对目的信息进行归类,然后进行分析[5];需要得到用户的页面访问习惯时,通常采用频率分析方法;如果想要得知用户所访问的页面间关系,例如访问某一产品页要根据前一页的产品介绍或者其他用户的评论,此时往往使用信息关联模式分析法;当需要对访问行为进行归类,以得到页面的访问信息时,往往要进行特征聚类分析。网络日志通过以上挖掘方法,能够将所需的信息以较为清晰的形式展示出来,得到有价值的信息,例如可以通过页面间的关联分析,得到自动的页面导航;通过使用频率分析方法,能够客观地评价出页面的重要程度;通过对分类的分析,能够对用户倾向性进行判断,得到该页面是否有良好的用户体验等[6]。

1.2　目前面临的问题

web数据最基本的属性是互联网,因此挖掘的难点和突破口均是在网络方面。对互联网自身的情况进行分析,分布式的站点结构往往使得数据存储的位置在地理上也是分散的,多源用户结构导致其用户多并且具有较强的动态性[7,8],网络异构的特性使得数据也往往呈现异构的特性,这些导致了数据挖掘面临着许多现实的问题:一方面是数据计算性能和网络传输性能的不匹配,另一方面是如何对网络日志进行有效获取、整合和处理[9,10]。因此,首先需要对分布式的网络日志结构进行深入分析,抽象更加适用的网络架构模型,其次需要对数据挖掘算法进行讨论,从而得到更高性能并且满足并行分布式的方法,从而保障挖掘的效率。本文着重讨论了高性能计算的相关策略。

2　改进型迭代算法

2.1　融合MapReduce思想

MapReduce是一种高效的面向分布式系统的编程模型,能够对大规模数据集进行良好的处理支持。MapReduce最重要的部分是任务调度,能够将一个任务进行细分,粒度更细的任务根据处理节点的性能,选择合适的节点进行处理,这种机制使得处理效率高的节点担负更多的任务,通过抑制低效节点来提高整个系统的效率。

数据键/值构成了MapReduce的基本结构,并通过映射和化简2个操作最终完成整个操作。映射实际上是一个分拆过程,完成了将输入数据大量的拆分,最终成为小的信息分段,这些信息分段最终交给每一个计算机处理单位进行处理,此时的计算模式即是分布式计算,然而化简这个过程又可以将所有拆分的数据进行合并,形成最后的结果。其包括2个重要的节点,分别是负责数据处理的worker和用于任务调度、处理共享数据的Master。这2个处理节点执行文件输入、worker分配切割文件、本地文件处理、文件合并操作、结果输出等过程。具体执行以下步骤:

1）MapReduce将文件切割成大小不等的份数,不同的处理机对各个备份进行处理;

2）由Master节点配置空闲worker节点,并且进行子任务分配,包括Map子任务和Reduce子任务;

3）分割文件会被worker节点读入,由该节点处理生成最后所需的键值对,Map处理最终的中间结果,并写入本地的硬盘中;

4）分区函数将所有的中间数据进行分区,Master根据本地硬盘的位置,发送给Reduce子任务所在节点进行处理;

5）Reduce执行节点任务,通过本地硬盘所保存的Map节点信息,进行中间信息key排序,合并相同的key;

6）对Reduce节点遍历执行后的中间信息进行排序,并向用户进行Reduce函数传递,最终得到输出文件;

7）所有任务完成后,Master节点能够将所有的结果信息输入至用户处理,相应程序再对其进行组合,即为所需输出。通过对MapReduce过程的分析可知,这种映射化简的思想非常适合于网络数据,能够充分利用有限的网络带宽。这是因为整个过程是在各个子节点完成的,并且本地化的处理使得在程序运算时数据并没有进行传输,最终向Master进行结果传输时才占用带宽资源,因此有利于降低带宽压力和分担计算压力,大大降低数据传输时间,提高整个过程的效率。

2.2　融合多服务器并行算法思想

面对着呈指数级增长的网络信息,web挖掘面临着数据计算和网络传输的双重压力。挖掘算法研究尤为重要,例如数据结构不同,使得算法对异构信息的输入和输出有不同的考虑;网络数据序列信息结构是不同的,因此信息量越大会使得序列越不同,存储这些信息又会有更多的问题;对数据库扫描次数的考虑,数据量越多,整个扫描量就会增加。

多服务器并行算法的基本思路是,将所有的计算过程分配到分布式服务器上,而不是远程服务器或本地的计算机,通过互联网,所有计算进程以服务的方式对用户需求进行支持。从该过程可以看出,多服务器并行算法思想对资源的调动和分配是动态的、可伸缩的,能够建立在互联网基础上的海量数据和资源知识发现。

网络传输速度与进程计算速度相差很大,因此尽量减少传输量是解决网络资源拥挤的关键。计算与存储整合的思想对网络系统的数据处理有关键的作用,将输入数据在集群机器本地磁盘上进行保存,能够大大降低传输资源带来的开销。通过将数据文件进行划分,把每个块在不同的机器上保存副本。根据Master记录器的位置数据,业务执行节点执行任务。输入数据几乎均通过在本地机器读取,对网络资源的要求很小。

多服务器并行算法系统能够对计算与整合思想进行良好的支持,除此之外,数据文件的备份问题同样也需要进行深入考虑,计算和存储能够根据节点的状况进行移动和调整,从而可以同时进行计算和存储的迁移。采取副本存储的方法是计算迁移的基础,根据副本的相应算法能够找到合适的数据。由于网络传输的影响,系统通过计算进行迁移远远快于通过网络迁移。副本策略能够实现将计算迁移至副本所在地,从而高效地完成存储和计算过程。

2.3 算法设计

迭代算法是计算机解决问题的常用方法,充分利用了计算机能够无误重复高效执行的优势,适合在多种环境下运用。数据挖掘是面向计算的一种信息处理,其本质还是算法的运用,因此大量的算法开始在数据挖掘领域使用,例如聚类分析、关联分析等。各种算法有其各自的优势和特点,关联规则的方法通常能够对网络日志进行高效分析,也在特征信息挖掘上有优势。关联规则的挖掘通常包括2个步骤,一是频繁项集的查询,二是分析频繁项集得到关联规则。本文设计的迭代算法,目的即是找到这些频繁项集。

为了避免在迭代过程中出现冗余的候选项集,以及对数据库的重复查询,本文结合多服务器并行算法的思想,将挖掘工作分配到计算节点并行处理,通过对局部的项集整合,最终得到全局项集。整个挖掘算法过程如下。

1)确定规则的置信度最小值和支持度最小值。

2)申请空闲节点:确定挖掘请求,任务调动中心向节点域请求XML文件,通过其提供的节点空闲情况,得到服务节点的机器名、IP地址等信息。任务调动中心再将得到的服务节点信息发送给算法存储单元。

3)得到局部项集:服务节点对各个本地的数据库进行扫描,得到事物数目、项出现频率,然后通过下面算法得到局部候选项集1:

```
frequent=new find_frequent_1-itemsets();
gen=new apriori_gen();
L1=Frequent (D);
for(k=2;Lk-1≠Φ ;k++) {
Ck=gen(Lk-1, sup_min);
for each node t ∈ D{
Ct=subset(Ck,t);
for each candidate c ∈ Ct
c.count++;}Lk ={c ∈ Ckc.count≥sup_min} }
return L= ∪ k Lk;
```

其中,以k-itemset代表k维项目集;Lk代表具有最小支持度的最大项目集;Ck代表候选最大项目集。

4)局部项集算法进行迭代:步骤3)得到了局部候选项集1,将其发送至Master可以计算出全局项集1,再通过全局频繁项集1,发送到服务节点得到精度更高的局部频繁项集1,而局部项集2可以由局部项集1得到。再一次迭代执行挖掘流程及局部项集算法,扫描本地数据库,得到项的出现次数,新局部候选项集2及结果发送至Master。最终得到满足所需的频繁项集,并且该频繁项集的最小支持度符合要求,之后根据置信度阈值得到关联规则。

3 基于改进型迭代算法的web挖掘模型

利用上文提出的挖掘算法,设计了融合多服务器并行算法思想的web挖掘模型。用Master代表主控节点,Service代表服务节点。所有服务通过1个主控节点进行调度和管理计算进程,另外1个节点是计算存储节点,负责提供具体的挖掘方法。服务节点的主要功能即是将其管理的功能存储好XML文件,并且执行相应的处理。Master将根据服务节点的处理得到最终的结果。整个系统可以分为3层,分别是信息层、算法层和执行层。信息层获取挖掘需求,生成挖掘算法需求,算法层根据算法需求调取适用算法并传递给执行层,执行层进行挖掘得到结果并返回给信息层主控节点(图1)。

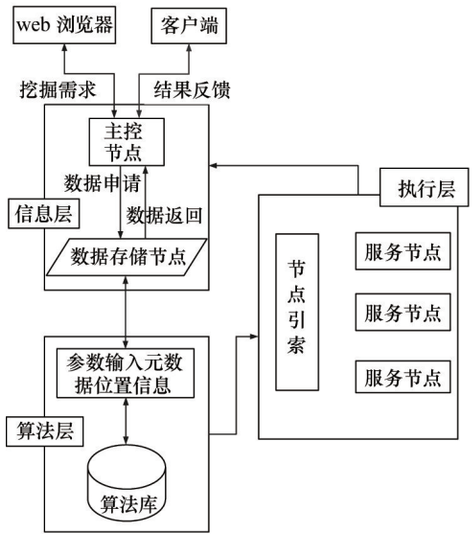

图1 基于改进型迭代算法的web挖掘模型

Fig. 1 Web mining model based on modified iterative algorithm

1）信息层。信息层的主要功能即是接收挖掘需求，调动整个数据挖掘过程，并且返回挖掘结果。

信息层提供了支持分布访问的数据接口，获取网络上的信息或者传递给节点信息。采用一定的机制将网络上的信息进行解析，生成对应的XML文件，例如网络日志文件，并且将生成的XML文件存储到系统中。该XML文件同时需要在数据节点上备份，从而避免因数据节点损坏导致的信息丢失。信息层对所有需求信息，例如用户浏览页面、购买基础数据等能够进行长期的存储，并且可以对分布式的数据集进行自动更新，支持数据节点的添加、删除和更新。XML文件是由一个主节点和多个子节点组成的，为了避免节点失效影响系统效率，本系统将XML进行备份，存于数据节点上。

数据存储节点能够对XML提供的基础数据进行存储，包括IP地址等信息，通常是通过分布式数据存储节点对XML文件操作和访问。数据存储节点另外一项功能是存储管理，用于控制数据节点的数据存放。数据节点是真实的数据存放点，并且实现了与客户端的数据通信。数据存储节点接收数据节点的工作报告，并且根据工作报告判断此数据节点是否正常，当出现异常时，需要由数据存储节点控制，将备份文件调出并再一次备份，这种机制提高了系统的可靠性。

一个XML文件由数据存储节点进行控制，以存储在分布式系统中。数据存储节点将会根据XML的大小进行分割，查询数据索引来得到空闲数据节点，这些分割部分存储至对应的空闲节点并进行备份。之后，客户端的操作将会通过数据节点直接进行。

2）算法层。算法层功能是提供挖掘算法，由于算法层独立处理，系统可以根据实际需要调整算法，或者对系统挖掘功能进行扩充。算法执行过程通常是，主控节点用来对挖掘需求进行分析，判断用哪种算法来执行，算法层提供该算法的存储，然后查询节点索引，得到在哪个节点上执行该算法，然后将该算法发送到原来数据的执行节点上。

3）执行层。执行层用于实际的业务调度、系统控制，所有的挖掘过程由主控节点统一调度。主控节点能够收集每个服务节点的工作状态，并且将空闲服务节点表随时进行维护，根据空闲服务节点表的情况，调用空闲服务节点。执行层接收客户端的挖掘请求，并且接收存储的数据信息，得到该用哪个挖掘算法执行，执行层向算法层申请具体的算法信息，然后算法层将此算法发送至相应的服务节点上。本地服务器直接执行计算过程，并将结果发送主控节点，主控节点将分布式服务信息进行汇总，并且返回客户端。

4　实验验证

为了验证此模型的性能，设计了一个模拟分布式系统。该系统包括1台主控节点服务器、1台算法层服务器、3台服务节点服务器、1台客户端服务器，并且相应服务器上安装Linux。为了更有效地模拟多服务器并行算法思路对模型的支持[9]，服务器也配备了Hadoop云计算系统。接下来对原始数据进行数据预处理，使用基于Hadoop+Hive的清洗工具对原始数据进行数据清洗。

为进行对比，将实验分成3组。

第1组，将主节点上的原始数据直接提供给算法层，算法层提供普通的挖掘迭代算法（例如Apriori算法），计算执行时间。

第2组，将原始数据文件复制成2份，分别保存至2台服务节点服务器上。使用本系统提供的挖掘算法，算法层将具体算法传输至服务节点，并行执行2个服务节点算法，得到执行时间。

第3组，将原始数据文件复制成3份，分别保存至3台服务节点服务器上。使用本系统提供的挖掘算法，算法层将具体算法传输至服务节点，并行执行3个服务节点算法，得到执行时间。

原始数据文件大小为50 GB左右，实验过程中对原有数据进行了采样，分别抽取出5、15、25 GB数据量的数据进行测试。

图2展示了3组实验的执行效率对比，并用执行时间的比例代表执行效率因子，效率因子用n表示，不同数据量用本系统模型测试所得结果用t_x表示，相应数据量原有系统测试所得结果用t表示，$n=10(t-t_x)/t$。执行效率因子越大，执行的速度越快。通过3组实验的对比可以得出，改进型迭代算法比传统的迭代算法效率更高，本地计算思想优于基于传输的计算思想；并且从第2、3组的对比看，增加服务节点后，随着被挖掘数据量的增加，整个算法执行时间有增长趋势。

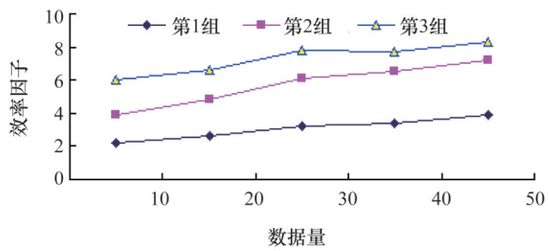

图2　3组实验执行效率对比

Fig. 2　Comparison of execution efficiencies of three groups

5　结论

针对网络信息的挖掘方式有很多种，需要根据不同的场景和需求制定合理的策略。与传统的数据挖掘相比，网络信息挖掘有更加特殊的要求。基于传统迭代方法原理，结合MapReduce和多服务器并行算法思维，提出了一种改进型迭代算法，并根据此算法提出了一种网络挖掘系统架构模型，力求能够对分布式网络系统的资源挖掘有更好的支持。实验初步证明了这种方法有更高的执行效率。由于该研究尚处在实验阶段，整个场景还仅限于实验模拟，当分布节点更多时，是否有其他干扰还不能确定，需要进一步实验和分析。

参考文献(References)

[1] 黄晓霞, 萧蕴诗. 数据挖掘集成技术研究[J]. 计算机应用研究, 2003, 20(4): 37-39.
Huang Xiaoxia, Xiao Yunshi. Research on the integration techniques of data mining[J]. Application Research of Computers, 2003, 20(4): 37-39.

[2] 李军怀, 周明全, 耿国华, 等. XML在异构数据集成中的应用研究[J]. 计算机应用, 2002, 22(9): 10-12.
LI Junhuai, Zhou Mingquan, Geng Guohua, et al. Research and application of heterogeneous dataintegration based XML[J]. Computer Applications, 2002, 22(9): 10-12.

[3] 程苗. 基于云计算的Web数据挖掘[J]. 计算机科学, 2011(增1): 146-149.
Cheng Miao. Web data mining based on cloud- computing[J]. Computer Science, 2011 (Suppl 1): 146-149.

[4] 胡开明, 陈建华. 一种改进的增量数据挖掘算法[J]. 计算机应用与软件, 2011, 28(8): 260-264.
Hu Kaiming, Chen Jianhua. An improved algorithm for incremental data mining[J]. Computer Applications and Software, 2011, 28(8): 260-264.

[5] 管忆军, 王勇, 何德牛. 一种采用函数迭代运算的数据流挖掘方法[J]. 广西民族大学学报, 2012, 18(1): 45-49.
Guan Yijun, Wang Yong, He Deniu. A data stream mining approach based on function iterative opration[J]. Journal of Guangxi University for Nationalities, 2012, 18(1): 45-49.

[6] 张浩, 景凤宣, 谢晓尧. 基于数据挖掘关联规则Apriori改进算法的入侵检测系统的研究[J]. 贵州师范大学学报: 自然科学版, 2011, 29(3): 84-87.
Zhang Hao, Jing Fengxuan, Xie Xiaoyao. The research of intrusion detection system based on improved apriori algorithm of data mining association rules[J]. Journal of Guizhou Normal University: Natural Sciences Edition, 2011, 29(3): 84-87.

[7] 彭宏玉, 柴旭光, 陈晓纪. 基于层次迭代思想的聚类算法的研究[J]. 唐山学院学报, 2011, 24(3): 86-87, 91.
Peng Hongyu, Chai Xuguang, Chen Xiaoji. The clustering algorithm of level lterated theory[J]. Journal of Tangshan College, 2011, 24(3): 86-87, 91.

[8] 赵洪英, 蔡乐才, 李先杰. 关联规则挖掘的Apriori算法综述[J]. 四川理工学院学报: 自然科学版, 2011, 24(1): 66-70.
Zhao Hongying, Cai Lecai, Li Xianjie. Overview of association rules apriori mining algorithm[J]. Journal of Sichuan University of Science & Engineering: Natural Science Edition, 2011, 24(1): 66-70.

[9] 柳莺, 赵艳红, 钱旭, 等. 数据仓库技术研究和应用探讨[J]. 计算机应用, 2001, 21(2): 46-48.
Liu Ying, Zhao Yanhong, Qian Xu, et al. Data warehouse technology research and application[J]. Computer Applications, 2001, 21(2): 46-48.

[10] 赵虎. 云计算环境下的关联数据挖掘算法实现[D]. 成都: 电子科技大学, 2011.
Zhao Hu. The implementation of association data mining algorithm In the environment of cloud computing[D]. Chengdu: University of Electronic Science Technology of China, 2011.

(责任编辑 王媛媛)

·学术动态·

中国科学技术协会

第37次中国科技论坛在上海召开

2014年12月6—7日,由中国科协主办、中国稀土学会承办,主题为“稀土资源开发与功能材料发展”的第37次中国科技论坛在上海召开。工业和信息化部稀土办公室副主任史瑞庭、华东理工大学原副校长卢冠忠、稀土材料国家工程研究中心副主任黄小卫、虔东稀土集团股份有限公司董事长龚斌担任本次科技论坛的学术召集人。来自政府部门、稀土科研院所、高等院校和生产企业的140余名代表与会。

中国工程院原副院长、中国稀土学会理事长干勇作“重大工程关键金属能源材料产业化技”报告,史瑞庭作“加大政策实施力度 推进稀土行业健康发展”报告,卢冠忠作“稀土催化材料的作用、创制和发展机遇”报告。

与会专家重点就稀土行业存在的问题、如何突破研发中遇到的瓶颈问题、稀土学科发展方向及促进稀土行业健康发展的举措等进行了深入交流。

详见中国科协网http://www.cast.org.cn/n35081/n35533/n38560/16138801.html。