文章编号:1005-8451(2005)02-0014-04

关联规则分布式算法的性能评价

陈 莉[1],罗学院[2]
(1.郑州铁路职业技术学院 信息工程系,郑州 450004;2.襄樊铁路分局 供电段,襄樊 441003)

摘 要:主要叙述两种基于WS Cluster WSs即工作站集群环境的分布式并行处理的有效性算法.第1种算法是在WSs间的关系数据比较小的表算法.另一种算法是对数据通信应用转换操作和对独立数据每接点进行大量搜索过程的简化,通过这些算法在WSs 中的实施,并对它们的性能作出评价。

关键词:CC 算法;Shift 算法;性能;关联规则;评价
中图分类号:TP311 **文献标识码**:A

Evaluation on distributed algorithmic performance of associated rule

CHEN Li[1], LUO Xue-yuan[2]
(1.Zhengzhou Railway Profession Techology Seminary, Zhengzhou 450004, China;
2.Water and Electricity Segment of Xiangfan Railway Subadministration, Xiangfan 441003, China)

Abstract: It was mostly described two new effective algorithms for parallel processing distributed in a WS Cluster environment. One was an algorithm in which the size of data transmitted between WSs was Smaller than that of the former algorithm. The other was an algorithm that reduced the number of scan processings at each node by dividing data and that used shift operations for data communication, and had implemented these algorithms on a WSs and had evaluated their performance.

Key words: count communication algorithms; Shift algorithms; performance; association rule; evaluation

本文叙述工作站集群的新分布式算法效应和性能的评价。CC 算法是利用广播操作在节点间进行小数量的数据传输。Shift算法通过划分数据和利用转换操作来减少在节点的扫描次数。在激光 SR2201 分布式存储并行计算机上及WSs 执行这些算法,比较通信代价和检测在通信搜索代价的多少。

1 关联规则的挖掘

1.1 关联规则

关联规则有多种,根据多种标准分类方法有:根据规则中所处理的值类型;根据规则中涉及的数据维;根据规则集所涉及的抽象层;根据关联挖掘的各种扩充.设I={i1,i2,…in}为一个数据项.设D={t1,t2,…tm}为项集记录,项表的每个记录要一致,如I 包含T。每个项包含记录X,如果I 包含X,那么S=Support(x)。在记录D的设置中,如果记录D的s% 中包含x。每个规则有两个测试定律和定义。这个推理规则X=>Y就是Support(x∪y),C的定义规则是X=>Y,在记录D 项中的意思是% 的记录D 中包含X 也包含Y,定义公式就是:

$$confidence\ X\Rightarrow Y = (Support(X\cup Y)/Support(x))\times 100 \quad (1)$$

1.2 Apriori算法

Apriori 算法使用了如下约定:(1)L_k:k维频繁项目集的集合,该集合中的每个元素包含量的部分为项目集本身和项目集的支持度。(2)C_k: k维侯选项目集的集合,是k 维频繁项目集集合的超集,也就是潜在的频繁项目集集合,该集合中的每个元素也包含项目集本身和项目集的支持度两部分。(3)任何项目集的元素都按某个标准(例如字典顺序)进行排序。(4)包含k个项目(k个项目为:C[1],C[2],…,C[k])项目集c用如下形式表示:C[1]·C[2]·…·C[k],由于c已经排序,所以排序准则有:C[1]<C[2]<…<C[k]。(5)如果c=X·Y,Y是一个m维项目集,也称Y是X的m维扩展(m-extension),相应地X为(k-m)维。

2 并行算法

2.1 基于Clusters的并行分布式系统的算法

从关联规则算法工具来处理候选项目集合中挖

收稿日期:2004-10-15
作者简介:陈 莉(讲师);罗学院(助理工程师)。

掘出大项目集合，必须提出大的频繁项目集合，因此这种算法代价很高。当实行并行处理时，通信代价也很高。NPA 并行算法减少了通信时间，主要是通过在每个节点上复制大项目集合。然后在每个节点执行相同操作并产生本地的频繁项目集。

这些算法适于在高速通信处理的并行计算机上应用，它们在 PCs 间或者通过以太网连接 WSs 执行。研究这两种算法是为了降低通信代价和基于Clusters环境上的性能评价。

2.1.1 CC算法

CC 算法不是在所有的节点传播大项目集合而是广播全局数量较小支持度的计数。每个节点的最大项目集集合是否满足特殊客户最小支持度值。在 k 项应用 CC 算法处理步骤如下：

（1）利用大的（k-1）- 项目集 L_{k-1}，通过 k-1 遍扫描产生候选 k- 项目集合 C_k，在每个节点进行复制。

（2）利用本地交易集合，计算出每个节点的候选 k- 项目集集合的支持度数 N^j_k 的值。

（3）在每个节点进行本地节点的支持度进行搜集，归并；全局计数 N_k 在其它所有节点进行广播。

（4）检查每个节点的候选项目集集合 C_k 是否满足特殊客户的最小支持度值。

（5）如果大 k- 项目集 L_k 为空，算法终止。否则，k=k+1，返回1继续进行。

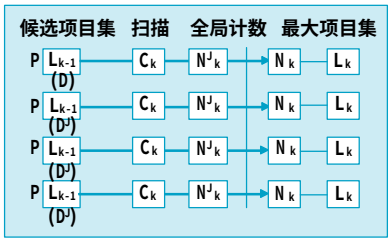

图 1　CC 算法图解

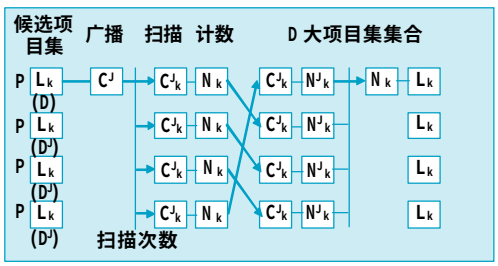

图2　Shift算法图解

2.1.2 Shift算法

当传输事务数据库时，Shift算法不能执行传送操作而是立即进行数据转换操作。运用Shift算法在第 K 遍进行处理步骤如下：

（1）利用大的（k-1）- 项目集 L_{k-1}，通过 k-1 遍扫描产生候选 k- 项目集合 C^j_k。候选 k- 项目集合 C_k 在每个节点进行划分。

（2）利用本地交易数量的候选 k- 项目集合在每个节点计算出现的支持度频率 N^j_k，在下个节点转换交易数据库 D^j。

（3）扫描所有的交易数据后，每个节点自己决定候选项目集集合 C^j_k 是否满足特殊客户的最小支持度。

（4）如果最大 k 数据项集 L_k 为空，算法结束。否则，k=k+1，最大数据项集在每个节点都继续执行。

2.2 代价分析

2.2.1 通信代价分析

通信代价就是数据信息通过关联规则的节点间传输的花费。

NPA 算法：当进行归并运算和大项目集集合广播。K 遍的通信代价 M_k^{NPA} 算法公式如下：

$$M_k^{NPA} = LAR_k \times (p-1) + n_k \quad (1)$$

$$M_k^{CC} = n_k \times (P-1) + n_k = n_k \times p \quad (2)$$

CC 算法：这个算法仅是广播计数，k 遍的通信代价 M_k^{CC} 的算法为公式（2）。Shift 算法：当转换所有交易数据和进行归并运算时，这个算法传输的数据，k 遍的通信价 M_k^{Shift} 算法为公式（3）。

$$M_k^{Shift} = D \times (P-1) + n_k \quad (3)$$

由此（1）（2）（3）公式的大项目集集合或者关系数据的大小比较，可以得出通信价 M_k^{CC} 比 M_k^{NPA} 和 M_K^{Shift} 的值都小。

2.2.2 搜索代价分析

搜索代价就是通过关联规则在所有的节点对全部数据项进行一次搜索的代价。并行算法利用大项目集集合产生候选项目集集合。搜索交易数据库中候选项目集集合，计算出候选项目集集合的数量。

NPA 算法：采用这个算法，在每个节点算出候选项目集集合，每个节点产生大项目集集合。在 k 遍的搜索代价 S_k^{NPA} 为公式（1）。

CC 算法：采用这个算法，在每个节点计算出候选项目集集合，全部节点产生大项目集集合。K 遍的搜索代价 S_k^{CC} 为公式（2）。

$$S_k^{NPA} = CAN_k \times P + LAR_k \quad (1)$$

$$S_k^{CC} = CAN_k \times P + LAR_k \times P = (CAN_k + LAR_k) \times P \quad (2)$$

Shift算法：采用这个算法，在每个节点算出候选项目集集合，每个节点产生大项目集集合。K遍的搜索代价S_k^{Shift}为公式（3）：

$$S_k^{Shift} = (CAN_k \div P) \times P + LAR_k = CAN_k + LAR_k \quad (3)$$

搜索代价S_k^{Shift}比S_k^{CC}和S_k^{NPA}的值都小。为用CC算法，每个节点都计算出候选项目集集合，全部节点产生大项目集集合，因此S_k^{CC}比S_k^{NPA}或者S_k^{Shift}任何一个都小。每个节点执行相同的搜索处理，然而搜索代价不能说明每个算法的处理时间。

3 实验

3.1 实验环境

主要在WSs和激光并行计算机上对NPA算法，CC算法和Shift算法进行测试。

WSs：WSs是由连接与100 mps的以太网（via）的32倍500 Hz的阿帕器21164WSs组成的。

激光并行计算机：激光并行计算机是分布存储的并行机，每个节点的连接都通过高速的三维网关。采用16位的激光机。

3.2 编程环境

通过基本的C语言程序MPI对算法编程执行。MPI就是为有效而安全地使程序能在许多平台上运行。

3.3 并行算法的执行

把所有的数据存放在数据仓库内，为统计分析和数据挖掘提供了更灵活的方式。不仅可以用Web分析工具提供的方法得到网站运行状况的报告，有经验的高级用户，可以直接用其他的数据访问方法（如OLAP）访问数据，或按照自己的爱好定制特定形式的报告。由于数据仓库本身就是为查询和数据挖掘所设计，因此采用数据仓库可以得到更高的效率和具有开放式的体系结构。采用关系数据库组成的1 000个关系，关系的平均长度是5，项目数是50。由于POS ID的相似处，自然数作为关系项。这个数据库除以每个节点。

实验测试报告，设关系数据为变量，以及最少维护和节点数。测试进行10次而得到一对确定的关系项数、最小的维护数和节点数的值，并且报告出最少的测试时间。

4 性能评价

4.1 执行时间和最小支持度之间的关系

3种算法的执行时间在最小支持度绘制出当节点数是8，交易项目集的集合数是5 000 h的结果。当最小维护为小时，Shift算法比NPA算法和CC算法好，但是也显示出NPA算法和Shift算法的执行次数快速增加。另一方面，Shift算法的执行时间不会因执行次数的增加而增多，但是，NPA算法和CC算法却会增多。

4.2 执行时间和交易数间的关系

所有算法的执行时间和通信时间的增长是和关系项数成比例的。Shift算法的指令执行时间比其它算法都短。当交易项目集的集合数大于2 500 h，并行性条件就非常明显了。图3说明当关系项数增加时，并行算法和序列算法的执行时间是成比例的。

4.3 执行时间和节点数间的关系

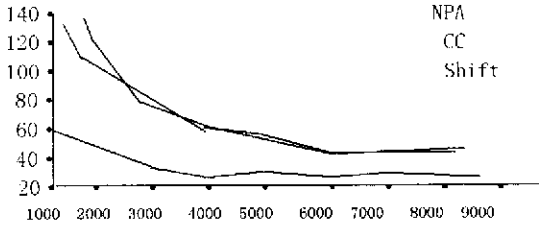

图3 并行算法与序列算法的执行时间的比例关系

4.4 在并行计算机上运行的结果

NPA算法和CC算法的通信次数在并行计算机上运行比在WSs条件下运行的少。另一方面，Shift算法的通信时间在WSs环境下运行比在激光机上运行少。

5 讨论

5.1 在工作站集群中的执行

对NPA算法和CC算法的执行次数随着最小支持度的减少而快速增加。这是因为当最小支持度的减少，候选项目集集合数在增加，在交易事务数据库中这些算法依着每个节点扫描候选项目集。然而因为Shift算法的公式是候选项目集集合除以每个节点，执行的扫描时间比其它算法的执行时间短。当最小支持度的减少时也不会快速增加。

价值分析说明CC算法的通信价最小，但是NPA算法和Shift算法的执行时间的测试是相同的。这好像是与测试中用的交易事务多少有关系。

5.2 在并行计算机上执行

文章编号:1005-8451(2005)02-0017-04

钢轨伤损管理信息系统的设计与实现

张树艳,郭年根,吕春英
(铁道部 信息技术中心,北京 100844)

摘 要:介绍钢轨伤损管理信息系统的设计与实现,包括项目的开发背景、系统功能框架、系统体系结构,描述了具体功能的实现方法。

关键词:钢轨伤损;铁路工务管理;信息系统;统一建模语言
中图分类号:U213.4;TP39 **文献标识码**:A

Design and implementation for Rail Defect Management Information System

ZHANG Shu-yan, GUO Nian-gen, LÜ Chun-ying
(Information Technology Center, Ministry of Railways, Beijing 100844, China)

Abstract: It was introduced the design and implementation for Rail Defect Management Information System, discussed the development background, framework of system function, system architecture, and described the implementation of all the function.

Key words: rail defect; permanent way management; Information System; UML

利用先进的信息技术和手段收集工务部门通过钢轨探伤车、探伤仪以及人工检测到的钢轨伤损信息(包括断轨信息),统计和分析钢轨伤损情况,从而及时有效地处理和监控伤损钢轨,确保铁路行车安全已成为当务之急。准确详尽地掌握钢轨伤损信息能够为工务部门制定钢轨探伤计划、大中维修计划和钢轨使用战略提供重要依据,从而合理利用建设资金并有效提高工作效率和管理水平。

收稿日期:2004-08-23
作者简介:张树艳,工程师;郭年根,高级工程师。

钢轨伤损管理信息系统由运输局基础部组织,铁道部信息技术中心开发,是铁路工务管理信息系

所有算法的执行次数在并行计算机上都比在WSs环境下执行的次数多。从而认为,这是因为并行计算机可用的CPU存储空间不足。另一方面,在并行计算机上,通信时间是非常稳定的,节点间的通信时间也比较少。换句话讲:在WSs环境下,通信时间和执行时间是成比例的,通信时间对执行时间有很大的影响。在WSs环境下,Shift算法的通信时间在使用转换操作比CC算法的时间少,就是传播操作。另一方面,在并行计算机上Shift算法的通信时间比通信量算法多。Shift算法在WSs环境下能有效地执行快速转换操作。

6 结束语

关联规则的定义及术语和关联规则挖掘的几种算法;在Clusters计算机环境下用分布式算法对交易事务数据库进行并行处理而找出有效的关联规则。在WSs环境下和在并行计算机上进行这些算法的实验,从而对它们的性能给出评价。当候选项集的集合数合在WSs环境下节点数较多时,发现Shift算法是效率最好的。这是因为在WSs环境下,通信时间和执行时间是成比例的。

参考文献:

[1] 陈 栋,徐洁磐. Knight:一个通用知识挖掘工具[M]. 计算机研究与发展,1998,4():338-343.
[2] 朱扬勇,周 欣,施伯乐. 规则型数据采掘工具集AMINER[M]. 高技术通讯,2000,10,3():19-22.
[3] 路松峰. 大型关联规则数据库开采算法[D]. 华中理工大学博士学位论文,2001.