

关联规则的并行挖掘模型研究

袁 平 张 伟

(重庆教育学院计算机与现代教育技术系 重庆400067)

摘 要 针对单机数据挖掘算法的现状,本文详细分析了 Agrawal 关联规则的并行挖掘模型,提出了基于客户机/服务器的并行挖掘模型和算法,建立了原型系统,该模型可以有效地用在关联规则的并行挖掘过程中,它能在一定程度上克服数据传输量大,无法及时得到所需数据而导致的效率不高的问题^[1]。

关键词 数据挖掘,关联规则,并行,客户机/服务器

Research on Parallel Mining Model of Association Rules

YUAN Ping ZHANG Wei

(Department of Computer & Modern Education Technology, Chongqing Education College, Chongqing 400067)

Abstract With the present situation of one machine data mining algorithms, the parallel mining model of Agrawal association rules is analyzed detailedly in this paper, a client/server based parallel mining model and algorithm are proposed, and a prototype system is created in the meantime. This model can be efficiently used in the process of parallel mining association rules, it can partly overcome large data transmission and low efficient problem caused by not receiving the needed data in time.

Keywords Data mining, Association rules, Parallel, Client/server

1 引言

在研究高效的单机算法的同时,多机并行、分布式的系统结构对于提高挖掘海量数据的速度和质量起到决定性的作用,本文分析了 Agrawal 关联规则并行挖掘算法的不足,提出了基于客户机/服务器模型的数据挖掘模型,在一定程度上弥补了 Agrawal 的不足。

定义1(支持度) 若数据库 D 中, $s\%$ 的事务包含 $X \cup Y$, 则关联规则 $X \rightarrow Y$ 的支持度为 $s\%$ 。

定义2(置信度) 若包含项目集 X 的事务有 $c\%$ 的也包含项目集 Y , 则关联规则 $X \rightarrow Y$ 的置信度为 $c\%$ 。

Agrawal 的关联规则发现包括两个问题^[5~7]: (1) 寻找满足最小支持度的频繁项目集; (2) 用频繁项目集产生规则, 例如: $ABCD$, AB 是频繁项目集, 如果 $AB \rightarrow CD$, 且通过计算置信度 $conf = support(ABCD) / support(AB)$, 如果 $conf$ 大于最小置信度, 则规则 $AB \rightarrow CD$ 就成立。它代表的意思是: 顾客购买了商品 AB 的同时, 也购买了 CD 。

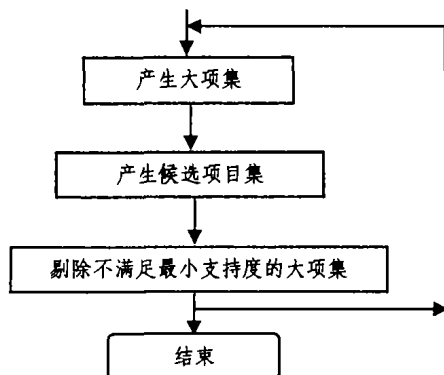


图1 寻找满足最小支持度频繁项目集的流程

问题(1)的解决思想是: ① 首先生成长度为1的大项集, 记为 $L[1]$; ② 在 $L[k]$ 的基础上生成候选项目集 $C[k+1]$; ③

用事务数据库 D 中的事务对 $C[k+1]$ 进行支持度比较以生成长度为 $k+1$ 的常用项目集 $L[k+1]$, 计算每个候选项目的支持度, 如果大于最小支持度, 则加入到 $L[k+1]$ 中; ④ 如果 $L[k+1]$ 为空集, 则结束, $L[1] \cup L[2] \cup L[3] \dots$ 即为结果。

问题(1)的解决流程如图1所示。具体的算法如下:

```

(i)  $L[1] = \{\text{长度为1的项目集}\}$ 
(ii) for  $(k=2; L[k-1] \neq \emptyset; k++)$  do
begin
(iii)  $C[k] = \text{apriori-gen}(L[k-1])$  // 新候选项目
(iv) for all transactions  $t \in D$  do begin
 $C = \text{subset}(C[k], t)$  //  $t$  中的候选项目
For all candidates  $c \in C$  do
 $c.count++$ ;
end;
 $L[k] = \{c \in C[k] | c.count \geq \text{minsupport}\}$ ;
End;
Answer =  $L[1] \cup L[2] \cup L[3] \cup \dots$ 

```

下面是候选项目集的产生算法:

```

Apriori-gen( $L[k]$ )
(i) Insert into  $C[k]$ 
Select  $p.item-1, p.item-2, \dots, p.item-(k-1), q.item-(k-1)$ 
From  $L[k-1]p, L[k-1]q$ 
Where  $p.item-1 = q.item-1$ 
 $p.item-(k-2) = q.item-(k-2)$ 
 $p.item-(k-1) = q.item-(k-1)$ 
(ii) for all itemsets  $c \in C[k]$  do
for all  $(k-1)$ -subsets  $s$  of  $c$  do
if  $(s \notin L[k-1])$  then delete  $c$  from  $C[k]$ 

```

例如:

$L[3] = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{1, 3, 5\}, \{2, 3, 4\}\}$

得到 $C[4] = \{\{1, 2, 3, 4\}, \{1, 3, 4, 5\}\}$

经过(ii)后得到 $C[4] = \{\{1, 2, 3, 4\}\}$, 因为 $\{1, 3, 4, 5\}$ 有子集 $\{1, 4, 5\}$ 不在 $L[3]$ 中。

2 Agrawal 关联规则并行挖掘模型分析^[2~4]

2.1 支持度分布计算模型

该算法解决的是关联规则的并行挖掘问题, 以前文提到的 Apriori 算法为基础, 采用的策略是不共享设备, 各处理机拥有各自的处理器和存储设备, 处理机通过网络连接。其模型

如图2所示。

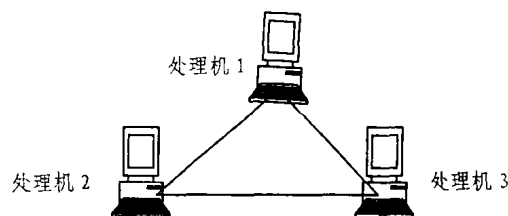


图2 支持度分布计算模型

具体步骤:

当 $k=1$ 时,各处理机利用各自存储的数据生成局部的 C_1 ,由于各自的数据不同,因此,必须交换本地的 C_1 得到统一的 C_1 。

当 $k>1$ 时

(1)每个处理机 P' 利用各自存储的 L_{k-1} ,通过连接,删减,得到本机上的所有 C_k 。

(2)处理机 P' 扫描它的数据 D' ,得到 C_k 的本地支持度。

(3)处理机 P' 同步,彼此交流局部 C_k 的支持度,得到全局 C_k 的支持度。

(4)每个处理机 P' ,将最小支持度与全局支持度相比较,剔除 C_k 中不满足要求的集合,生成 L_k 。

每个处理机根据条件独自决定是终止或接着进行(1)。

在第一步,处理机 P' 根据各自存储空间的数据,动态产生局部候选项目集 C_1 ,因此,各处理机对候选项目集的计算是不相同的,所以,彼此交流局部的支持度就非常必要,文[5]中是通过在每个处理机上建立缓冲区来进行交换的。

2.2 数据分布计算模型

处理机的连接结构也是采用图2的结构。它的思想步骤是:

(1)处理机 p' 从 L_{k-1} 通过连接操作产生 C_k , p' 只对候选项目集 C_k ,其长度为 C_k 的 $1/n$,究竟怎样对应,就要根据机器的编号。各 C_k 连接起来就是 C_k ,但它们存储在不同的机器上。

(2) p' 利用本地的数据和接收到的从其它处理机发送来的数据计算项目集的支持度。

(3) p' 用最小支持度来衡量 C_k ,剔除不满足要求的集合,得到 L_k 。这时,各 L_k 分散在各处理机上,它们连接起来就成为 L_k 。

(4)处理机之间交流 L_k ,通过整理,合并,得到完整的 L_k ,为下一步产生 C_{k+1} 作准备,这时要求处理机同步,取得了 L_k 后,各处理机可以决定是进行下一轮,还是终止计算。

从上述分析可以看出:第一点,第(1)、(3)是显然的,而在第(2)步中,处理机在接受其它处理机通过数据页传来的数据的时候,它又要将自己的数据向其它处理机广播。第(4)步交流 L_k 与 C_k 类似。第二点,第(2)中异步传输各数据,导致大量多次传输,拷贝数据。除了对网络传输和通信要求很高以外,还要花费大量的时间在数据传输上。由于各处理机之间数据的强烈依赖,使得处理机之间很可能因不能得到需要的数据而空等待,以至于效率不高。

2.3 支持度分布计算模型与数据分布计算模型比较

(1)前者处理机之间交流的是 C_k 的支持度,后者交流是 L_k 。

(2)前者 C_k 的最大容量为 $|M|$,后者的最大容量为 $n \times |M|$ 。

(3)前者各处理机上处理的 C_k 是相同的,后者处理的 C_k 是不相同的;

(4)前者得到了完整的 C_k ,表现为一个可见的集合,后者 C_k 只有通过连接才能得到 C_k , C_k 存储在不同的机器上。

(5)前者网络传输少,后者需要大量的传输。

3 基于客户机/服务器的并行挖掘模型和算法

针对上述模型,本文提出了基于客户机/服务器的并行挖掘模型和算法,以图解决数据量传输巨大,以及容量受到限制的问题。

3.1 算法及模型

算法首先通过 partition 算法将数据分配给 n 台不同的处理机,经过各处理机的计算,相互传递数据,最后计算得到结果。其算法如下:

server process //服务器端算法

i)与 client 建立联系。

ii)将数据分为 n 部分,并分配给 n 个 client。

iii)server 等待并接收来自 client 的局部大项集,计算产生全局候选项目集。

iv)server 对全局候选项目集进行剪枝,并将剩下的候选项目集分配给 client。

v)client 产生新的候选项目集,计算其本地支持度返回给 server。

vi)server 接收到候选大项集的各局部支持度,计算出全局支持度。

vii)产生最终大项集。

client process //客户端算法

i)从 server 处获得数据。

ii)产生局部大项集。

iii)将局部大项集发送给 server。

iv)接收全局候选项目集。

v)对接收到的未包含于自身项目集中的项目,计算其本地支持度。

vi)发送本地支持度给 server。

并行运算结构模型如图3所示。

3.2 分析

该模型和算法也是基于交易数据库,server 首先将数据分配到 client,然后等待,client 通过计算,将本地每一遍运算过程产生的大项集传送到 server,在 server 形成全局候选项目集,通过对候选项目集进行剪枝,然后把剩余候选项目集送到 client,client 产生新的候选项目集,并计算它本地的支持度,把支持度返回给 server,server 接收到候选大项集的各局部支持度,计算它的全局支持度,不断重复上述过程,直到产生最终的大项集。

该模型与 Agrawal 模型最大的区别是,处理机之间传输数据量少,处理机直接和服务器进行数据交换,避免了处理机之间由于数据的强烈依赖而出现效率不高,另外该模型较 Agrawal 模型简单,可靠性高。

结束语 传统的单机串行算法已经不能适应快速、实时的知识需求,研究面向多机、并行、分布式的数据挖掘模型越来越重要,使得多机并行数据挖掘具有很高的实际价值,本文的模型很容易在局域网实现,通过该模型,对大容量数据的处理相对容易,可以提高数据挖掘的精度,以及数据挖掘的质量。

(下转第205页)

据缓冲器。ALTERA 公司的 EP1K100QC208 芯片内部包含 12 个 EABs, 每个 EAB 最多能配置成 4096 比特的存储器, 而且只能配置成一个存储器, 就是两个或两个以上的存储器不能公用一个 EAB, 当一个 EAB 的部分空间已经被占用, 剩下的空间也不能重新分配出来。当图像进行小波变换时, 每一级都要使用 6 个 FIFO 存储器, 共占用 6 个 EABs, 这样造成了片内存储空间的巨大浪费, 导致片内存储空间只能进行两级小波变换。为了充分利用片内的存储空间, 可以将片内所有 EABs 都分配给一个双口存储器, 小波变换的 FIFO 操作由两个存储器控制器来完成, 各级小波变换共用一个大容量双端口存储器, 这样一片 EP1K100QC208 就能完成对大小为 256×256 的图像进行四级小波变换。

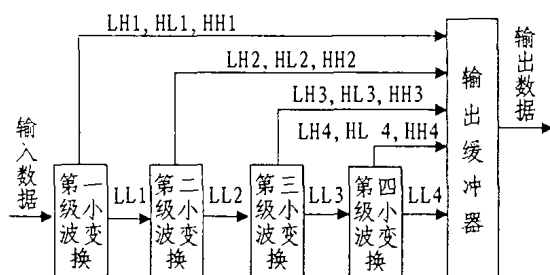


图5 DWT 的四级流水线结构

图6中, a 为一幅 256×256 的夜间红外成像照片, b 为采用本文小波变换器处理后的小波域图像。当采用 40MHz 的时钟输入时, 每幅图像的变换时间约为 30ms, 完全能达到每秒 25 帧的要求。

结论 本文结合实时红外图像压缩系统的要求对 JPEG2000 推荐的 DWT 算法进行一系列的改进, 使小波变换的硬件开销大幅度减少, 并将该 DWT 处理器用 EP1K100QC208-3 进行了验证, 经过设计、仿真和下载, 各项

性能基本达到了预期的效果, 整个系统的资源消耗如下: ①使用逻辑单元 (LEs) 3148 个, 占 63%; ②使用嵌入式阵列块 (EABs) 12 个, 49152 位, 占 100%。

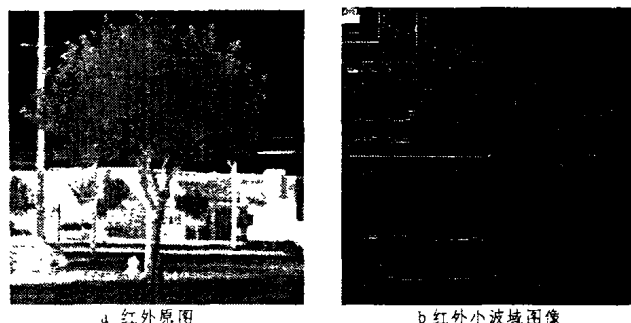


图6 红外图像的小波变换图

在设计过程中充分考虑了动态范围和截尾误差, 经过四级小波变换后的重构图像失真度小, 对图6的红外图像进行四级小波变换后的重构图像峰值信噪比 (PSNR) 为 237.3dB, 这种失真用肉眼是无法察觉的。

参考文献

- 肖山竹. 高速图像小波分解算法与 FPGA 实现. 电子技术与应用, 2003(4)
- 刘雷波, 等. JPEG2000 小波变换器的 VLSI 结构设计. 电子学报, 2002(11)
- A theory for multi-resolution signal decomposition: The wavelet representation. IEEE transactions on pattern analysis and machine intelligence, 1989, 11(7)
- Mallat. ISO/IEC 15444-1. Information technology-JPEG2000 image coding system. 2000
- 朱秀昌, 等. 数字图像处理于图像通信. 北京邮电大学出版社, 2002

(上接第182页)

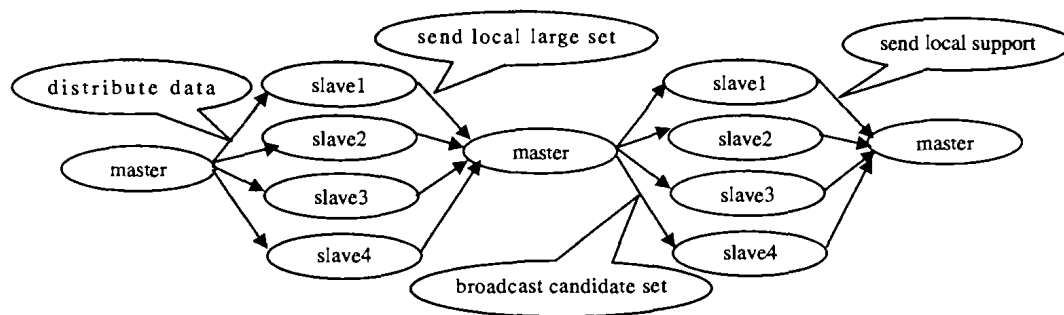


图3 并行运算结构模型

参考文献

- Chen Y. An Efficient Parallel Algorithm For Mining Association Rules In Large Database. In: Proc. of the 10th Canadian Conf. on Computation Geometry, 1998
- Agrawal R, Srikant R. Mining Sequential Patterns Research Report RJ 9910. IBM Almaden Research Center, San Jose California, Oct. 1994
- Agrawal R. Parallel Mining of Association Rules. IBM Almaden Research Center
- Oates T, Paul R, Cohen. Parallel and Distribute Search for Structure In Multivariate Time Series, Machine Learning ECML-97 Berlin New York
- Matthew D, Schmill, Oates T. A Distribute Approach To Finding Complex Dependencies In Data. Lederle Graduate Research Center Computer Science Technical Report 98-3
- Agrawal, Imielinski, et al. Mining Association Rules between Sets of Items in Large data. 1993
- Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 1996, 17(3): 37~54