一种分布式数据库关联规则挖掘算法①

曹文梁

(东莞职业技术学院 计算机工程系, 东莞 523808)

摘 要：现有的数据挖掘算法多是集中式环境下的数据挖掘处理，但目前的大型数据库多以分布式的形式存在，针对分布式数据挖掘算法 FDM 及其改进算法中存在的频繁项集丢失问题和网络通信开销过高的问题，提出了一种改进的基于关联规则的分布式数据挖掘算法 LTDM，LTDM 算法引入了映射标示数组机制，可以在保证频繁项集完整性的同时降低网络的通信开销。实验结果证明了算法的有效性。

关键词：分布式数据库；数据挖掘；关联规则；频繁项集；网络通信开销

Distributed Association Rules Mining Algorithm

CAO Wen-Liang

(Computer Engineering Department, Dongguan Polytechnic, Dongguan 543808, China)

Abstract: Most of the existing data mining algorithms are processing in the centralized systems; however, at present large database is usually distributed. Compared with the frequent itemsets lost and high communication traffic in distributed database conventional and improved algorithm FDM, an improved distributed data mining algorithm LTDM based on association roles is proposed. LTDM algorithm introduces the mapping indicated array mechanism to keep the integrity of frequent itemsets and decrease the communication traffic. The experimental results prove the efficiency of the proposed algorithm.

Key words:distributed database; data mining; association rules; frequent itemsets; communication traffic

现有的数据挖掘算法和模型大多是在集中式环境下基于数据库或数据仓库环境的数据挖掘处理[1]。随着计算机网络及分布式数据库的发展，大量数据资源存储在网络中的不同站点中，分布式的环境下的数据存储模式要求将数据汇聚在统一集中的数据仓库中，于是大量数据需要通过网络传输，使用有限的传输速率将海量数据汇聚在同一站点中，此时网络通信效率，信息安全和通信成本等促使人们在考虑数据挖掘算法时更注重不再将分散的数据汇聚于同一站点，而是提出能够适应分布式数据的分布式数据挖掘算法[2,3]。Agrawal，Imielinski 等基于分布式数据库首先提出了关联规则数据挖掘算法[4]，该算法针对无法进行集中处理的大量分布式数据库中的数据进行数据挖掘操作。

分布式数据库的快速发展和广泛应用对分布式数据挖掘算法的性能提出了更高的要求。已有的经典分布式挖掘算法在执行效率和网络通信开销方面已经不能很好的满足系统的需求[5]。本文在分析 FDM (Fast Distributed association rules Mining)算法[6]优缺点基础上，提出了一种基于改进的分布式挖掘算法 LTDM (Low Traffic Distributed Mining algorithm)，LTDM 算法可以在保证频繁项集完整性的前提下降低网络通信开销并提高算法的执行效率。

1 相关概念与定义

定义 1 设现存平行方式的分布式数据库系统 T，其中包含逻辑同构的数据库 n 部分：

$$T=\{T_1,T_2,T_3,\cdots T_n\}$$

① 收稿时间:2012-01-06;收到修改稿时间:2012-02-19

DB 为 T 的分布式数据库，站点 T_i 的数据库为 $DB_i(1 \le i \le n)$，$DB = DB_1 \cup DB_2 \cup \cdots \cup DB_n$

$$L = L_1 + L_2 + \cdots + L_n$$

其中 L_n 表示局部数据库 DB_n 的大小，L 表示全局数据库 DB 的大小。

定义 2 如果 $X.\sup(X.count/D)$ 和 $X.\sup^i(X.count^i/D_i)$ 分别表示 X 在 DB 和 DB_i 中的支持度，则称 $X.\sup$ 为 X 的全局支持度，$X.\sup^i$ 为 X 在站点 T_i 的局部支持度。

当 $X.\sup \ge \min\_\sup$ 时，可以判定 X 为全局频繁项目集；

当 $X.\sup^i \ge \min\_\sup$ 时，可以判定 X 为站点 T_i 的局部频繁项目集；

性质 1 当 X 为站点 T_i 上的局部(全局)频繁项目集时，可以判定 X 的所有非空子集均为站点 T_i 上的局部(全局)频繁项目集。若 X 不是站点 T_i 上的局部(全局)频繁项目集，则可以判定 X 的超集必定不是局部(全局)频繁项目集。

性质 2 当 X 是全局频繁项目集时，可以判定必定存在一个站点 $T_i(1 \le i \le n)$，使得 X 及其非空子集在站点 T_i 上是局部频繁的。

2 FDM算法简介

2.1 FDM 算法的实现

文献[6]在分析和总结分散数据集和集中式数据集相互间关系的基础上，提出了 FDM(Fast Distributed association rules Mining)算法，FDM 算法是一种快速的应用于分布式数据库系统中的关联规则数据挖掘算法，它的主要步骤分为以下 5 步：

(1) 生成本地候选项目集。在任意站点数据库系统 DB_i 上，基于本地全局频繁 k-1 项集 $Local\_G_{k-1}$，使用公式 C_{ik}= apriori_gen(LG_{k-1})，生成本地候选频繁项目集 C_{ik}。

(2) 本地剪枝操作。对于每一个本地数据集 $X \in C_{ik}$，对每个局部数据库 DB_i 进行扫描，根据扫描结果计算本地支持度，当 X 相对于站点 DB_i 是局部大的，则将 X 加入进 DB_i 的频繁 k-1 项目集 L_{ik} 中。

(3) 交换支持数操作。各站点发送支持数，并对全局频繁 k 项集的总支持数进行计算。

(4) 广播操作。各个局部数据库 DB_i 将各自的全局频繁 k 项集广播发送给其他站点上的局部数据库 DB_j。

(5) 对步骤(1)至(4)进行重复，当没有新的频繁项集产生时停止。

2.2 FDM 算法性能分析

FDM 算法的优点主要包括：(1)总结并利用了局部频繁集与全局频繁集之间的有价值关系，减少了消息传递量，降低了整体系统的通信开销；(2)每个独立的站点都可以选用局部剪枝操作和全局剪枝操作来对候选数据集进行裁剪处理。

在 FDM 算法中，利用了局部频繁集与全局频繁集之间的两个重要关系来对通信开销进行控制，以减少网络消息传输量。

关系 1：假设数据集 X 是全局大的，则必定存在一个站点 T_i (1≤i≤N)，数据集 X 及其所有子集在站点 T_i 是全局大的。

关系 2：若 CGi(k)=apriori_gen (LG(k-1))，则对任一 k>1，所有全局大的 k-项集 L(k)是 $CG(k)=\bigcup_{i=1}^{n} CGi(k)$ 的子集。

分析发现，在生成的候选数据集 CG(k)中，某些数据集在进行合计数交换操作之前就可以进行局部剪枝操作。在站点 T_i 的局部剪枝过程中，只用到了 DB_i 的局部支持合计数，其余站点也可以进行剪枝操作。在每次迭代完成时，可以得到候选数据集 X 的局部和全局支持合计数。在广播操作完成后，站点就可以应用接受到的广播消息在后续的迭代操作中进行全局剪枝操作。

从算法的流程中可以发现 FDM 的一些不足：(1)在 FDM 算法中，各个局部站点只是广播各自的支持数和局部频繁 k-项集 L_{ik}，而对不属于局部频繁项集的候选项集只进行简单的剪枝处理。(2)在本地剪枝时，FDM 算法扫描每一个局部数据库，计算本地支持度，仅当 X 是站点的局部大时，才加入到频繁 k-项目集中。总结可得，FDM 算法有一定几率在使用关联规则进行数据挖掘过程中造成频繁项集的丢失。

文献[7]中提出了一些改进方法：

(1) 在全局站点上将所有局部站点发送的频繁项集进行合并，以得出全局频繁项集的候选集。

(2) 全局站点将合并得出的候选数据集向所有局部站点广播，在局部站点上计算各自的局部支持数并确定本地的新增项目集。

(3) 局部站点将得出的新增项集信息和局部支持数信息发送到全局站点上去，全局站点对收到的数据

进行累加操作，得出各个全局候选数据集的全局支持合计数。

以上改进措施修正了 FDM 算法的缺陷，可以保证全局频繁项集的完整性，但是频繁的全局站点与局部站点的通信操作明显增大了网络的通信开销。

3 LTDM算法的实现

分布式数据库的关联规则挖掘算法的时间开销主要由两个方面决定：(1)频繁项集的确定；(2)网络的通信消息量。本文基于 FDM 算法以及其改进算法的不足，提出了一种新的改进算法——LTDM，LTDM 的思想是在全局站点和局部站点建立互相对应的频繁项集映射数组，使得各个局部站点和全局站点上的频繁项集数组可以实现一一对应，根据各个局部站点的挖掘结果，将对应的全局站点上的映射数组值置为 1，以简单的赋值操作代替整个频繁项集的传输工作，以降低分布式系统的通信开销。

LTDM 算法主要从以下方面对 FDM 算法进行改进：

(1) 在中心站点和局部站点上分别建立 n 个频繁项集的映射标示数组 B_i，中心站点的映射标示数组值为 B_0，局部站点依次增大。候选项集中的每一项在映射标示数组中都有对应位，在进行扫描操作后，当支持度满足 min_sup 条件时，数组对应位置为 1，不满足条件则保持为 0。

(2) 在中心站点上将数组 B_0 与局部站点发送来的映射标示数组 B_i 的各个位的值进行或运算，只要对应位上有一个值为 1，则数组 B_0 的值为 1，使用这个方法可以克服保证任一项集只要在任一局部站点上式频繁项集，数组 B_0 的值为一定 1，以避免 FDM 算法中忽略站点计数的问题，并保证 LTDM 算法频繁项集的完整性。

(3) 为避免全局频繁 k-1-项集在某个局部站点不频繁而引发的候选 k 项集缺失和频繁项集丢失问题，当每一轮的 k-频繁项集挖掘完成后，中心站点都进行 L_k 的广播操作。

(4) 各局部站点不再进行全局的广播，而是先将映射标示数组发送到全局站点进行标示，然后根据标示结果将频繁项集和支持数发送至全局站点，以降低网络的通信开销。

4 算法性能实验

为测试 LTDM 算法的性能，我们使用多台计算机组成了局域网以模拟分布式数据库系统。计算机的配置为：CPU AMD 速龙 3000+，内存为 1G，硬盘为 80G，网络为 100M 的以太网，子网段间使用 100M 的交换机相连，操作系统为 Windows Server 2003。本实验的数据来自于某大型连锁超市销售数据库中的 900000 个样本记录。算法是实验结果如图 1 和图 2 所示。

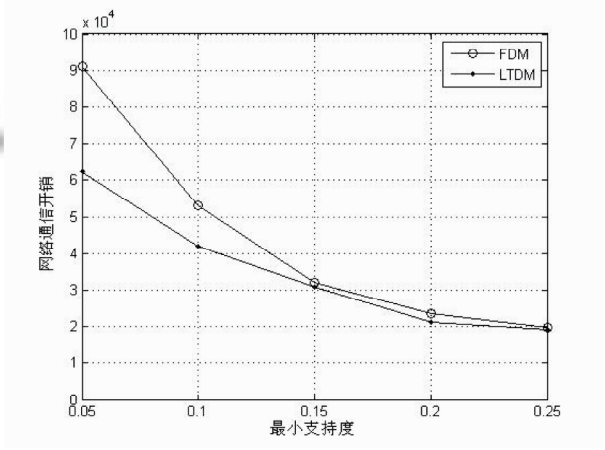

图 1　网络通信开销对比

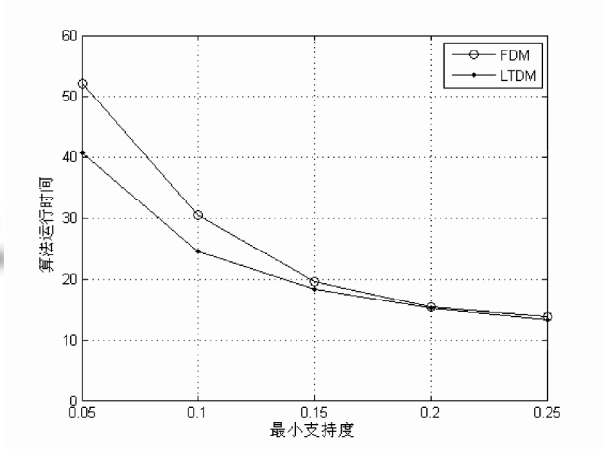

图 2　算法运行时间对比

图 1 为网络通信开销与最小支持度的关系，可以明显看出，LTDM 算法在网络开销上明显的少于 FDM 算法，在最小支持度为 0.05 时，LTDM 的网络开销仅为 FDM 算法的 67.61%。

图 2 为算法运行时间与最小支持度之间的关系，可以看到，LTDM 算法的执行效率在最小支持度高时与 FDM 算法基本相同，但在最小支持较低时，LTDM

算法的执行效率略优于 FDM 算法。

5 总结

分布式数据库系统中的关联规则挖掘算法的性能改进主要集中于降低网络通信开销和减少候选频繁项集上。传统的 FDM 算法容易造成频繁项集的丢失，而已有的改进措施又引发了网络通信开销过高的问题。本文在分析 FDM 及其改进算法的基础上，提出了一种基于改进的适用于分布式数据库系统中的关联规则挖掘算法 LTDM，LTDM 算法在全局站点和局部站点上引入了映射标示数组，可以在降低网络通信开销的同时保证频繁项集的完整性。实验结果证明，LTDM 算法在网络通信开销控制和算法运行效率方面具有比 FDM 算法更有优异的性能。

参考文献

1 Hahsler M, Bettina G, Hornik K, Buchta C. Introduction to arules a computational environment for mining association rules and frequent item sets,2010.

2 Novak PK, Lavra N. Webb GI. Supervised descriptive rule discovery:A unifying survey of contrast set, emerging pattern and subgroup mining.Journal of Machine Learning Research, 2009,10(09):43−48.

3 杨明,孙志挥,吉根林.快速挖掘全局频繁项目集.计算机研究与发展,2003,40(4):620−626.

4 Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases, ACM SIGMOD Conf, Management of Data,1993.

5 王春花,黄厚宽,李红莲.一种快速有效的分布式开采多层关联规则的算法.计算机研究与发展,2001,38(4):438−443.

6 Cheung DW, Han J, Ng V, Fu AWC, Fu Y. A Fast Distributed Algorithm for Mining Association Rules.Proc.1996 Int'l Conf. Parallel and Distributed Information Systems (PDIS '96).1996.31−42.

7 Nijssen S, Guns T, Raedt LD. Correlated itemset mining in roc space:a constraint programming approach. KDD,2009. 647−656.

(上接第 169 页)

量参数的不确定性，采用了神经网络自适应补偿控制对不确定参数进行补偿，并分别对补偿前后系统的跟踪性能进行了 MATLAB 仿真，通过对比结果表明，经过神经网络补偿后控制误差明显减小，从而证明了采用神经网络的反馈线性化非对称电液伺服系统方案的可行性。

参考文献

1 何玉彬,李新忠.神经网络控制技术及其应用.北京:科学工业出版社,2000.

2 王永超,金勇,王力,等.基于模糊神经网络的电液伺服系统建模.计算机仿真,2011,28(5):184−187.

3 刘长年.液压伺服系统优化设计理论.北京:冶金工业出版社,1989.

4 刘公信.不对称油缸电液伺服系统分析.煤矿机械,2008,29(8):77−79.

5 杜红彬,余昭旭.一类仿射非线性系统的自适应神经网络输出反馈变结构控制.控制理论与应用,2008,25(6):1042−1044.

6 倪敬,彭丽辉,项占琴,等.扩轧管电液伺服系统非线性建模与控制.机械工程学报,2009,45(5):250-255.

7 Chen ZX, Wang LY. Adaptable product configuration system based on neural network. International journal of production research,2009,47(18):5087−5107.

8 朱学彪,陈奎生,傅连东,等.基于 AMFC 的电液伺服系统控制算法研究.液压气动与密封,2008,28(4):57−6.