通信网告警加权关联规则挖掘算法的研究

李彤岩，肖海林，李兴明

(电子科技大学宽带光纤传输与通信网技术教育部重点实验室 成都 610054)

【摘要】关联规则挖掘算法是通信网告警相关性分析中的重要方法。在处理数量庞大的告警数据库时，算法的效率显得至关重要，而经典的FP-growth算法会产生大量的条件模式树，加权算法MINWAL(O)则需要多次扫描数据库，使得在通信网环境下挖掘关联规则的难度非常大。该文提出了一种高效的基于加权频繁模式树的通信网告警关联规则挖掘算法，算法性能测试表明，该算法与已有的加权关联规则挖掘算法相比较，节约了大量的存储空间，提高了算法的挖掘速度，对通信网的故障诊断和故障定位有着积极的意义。

关 键 词 告警相关性分析； 故障诊断； 故障定位； 加权关联规则； 加权频繁模式树

中图分类号 TN915.07 **文献标识码** A

Algorithm for Mining Weighted Alarm Association Rules in Telecommunication Networks

LI Tong-yan, XIAO Hai-lin, and LI Xing-ming

(Key Laboratory of Ministry of Education for Broadband Optical Fiber Transmission and Communication Networks, University of Electronics Science and Technology of China Chengdu 610054)

Abstract Mining association rules is one of the primary methods used in telecommunication alarm correlation analysis. The efficiency of the algorithms plays an important role in tackling with large datasets. A highly efficient algorithm of weighted association rules mining in telecommunication networks based on weighted frequent pattern tree is proposed. The performance test of the algorithm indicates that compared with other algorithms of weighted association rules mining, this one needs less memory and has higher temporal efficiency, which is significant for the network fault diagnosis and localization.

Key words alarm correlation analysis; fault diagnosis; fault localization; weighted association rules; weighted frequent pattern tree

通信网告警相关性分析是网络故障诊断的重要手段[1-2]，基于关联规则挖掘算法的告警相关性分析成为目前的研究热点。关联规则中最具代表性的是文献[3-5]提出的Apriori算法和类Apriori算法，以及文献[6]提出的FP-Growth算法。

通信网中告警信号的某些属性分为不同的级别，如严重告警、重大告警和一般告警等。不同QoS要求的业务对告警处理的程度也有所不同，应该给告警数据分配不同的权值，然后对加权的告警信息进行关联规则挖掘。加权关联规则挖掘算法的研究国内外已有了很多成果[7-8]。其中基于Apriori算法的一个共同缺点是需要多次的扫描数据库，严重地影响了算法的挖掘速度，如MINWAL(O)算法[9]；而基于FP-Tree的算法又会产生数量巨大的条件模式树，内存占用率非常高。

本文提出了一种高效的基于加权频繁模式树(weighted frequent pattern tree, WFP-tree)的关联规则挖掘算法，采用文献[10]提出的层次分析法为告警分配权值并克服多次扫描数据库和递归产生大量条件模式树的缺点，提高了算法的挖掘速度，节约了存储空间。最后通过仿真进行了算法的性能比较。

1 加权关联规则挖掘算法

1.1 加权关联规则相关概念

假设告警交易数据库为D，其项目的集合为$I=\{i_1, i_2,\cdots,i_n\}$，其中$i_j(j=1, 2,\cdots, n)$为预处理后的项集。告警交易数据库$D=\{T_1,T_2,\cdots,T_n\}$，其中每个交易$T_i$ ($i=1, 2, \cdots, n$)是项集I的子集。项集X的支持度计数

收稿日期：2007－06－18；修回日期：2007－12－21
基金项目：国家自然科学基金(60572091)
作者简介：李彤岩(1980－)，女，博士生，主要从事通信网告警相关性和数据挖掘方面的研究.

Support_Count(X)为D中包含项集X的交易数；X的支持度Support(X)= Support_Count(X)/T，T为D中交易的个数，$X\subset I$。当Support(X)≥minsup(minsup为最小支持度阈值)时，称项集X是频繁项集。

本文结合通信网告警的特点综合考虑了告警属性和网络拓扑信息，采用文献[10]提出的层次分析法为I中的每一个项目i_j分配一个权值w_j($0\leqslant w_j\leqslant 1$，$j=\{1,2,\cdots,n\}$)。项集$X$的加权支持度计数(weighted support count)为$\left(\sum_{i_j\in X} w_j\right)(\text{Support_Count}(X))$，$X$加权支持度为$\left(\sum_{i_j\in X} w_j\right)(\text{Support}(X))$。当$X$的加权支持度满足$\left(\sum_{i_j\in X} w_j\right)(\text{Support}(X))\geqslant W_{\text{minsup}}$($W_{\text{minsup}}$为预先给定的最小加权支持度阈值)时，称$X$是加权频繁项集。$X$的支持度计数应满足：

$$\text{Support\_Count}(X)\geqslant\frac{W_{\text{minsup}}T}{\sum_{i_j\in X} w_j} \tag{1}$$

式(1)给出了加权频繁项集的判断条件。参照文献[9]可知，令I为告警交易数据库D中所有项集的集合。设Y是一个q-项集，$q<k$。在剩余的项目集合(I-Y)中，假设前(k–q)个权值最大的项为$i_{r_1},i_{r_2},\cdots,i_{r_{k-q}}$，那么任意包含项集$Y$的$k$-项集的最大可能权值为：

$$W(Y,k)=\sum_{i_j\in Y} w_j+\sum_{j=1}^{k-q} w_{r_j} \tag{2}$$

式中 第一项是q-项集Y的权值之和；第二项是剩余的权值前(k–q)个最大值之和。结合式(1)和式(2)，可以推知如果包含Y的k-项集是频繁的，那么其最小支持数应为：

$$B(Y,k)=\left\lceil\frac{W_{\text{minsup}}T}{W(Y,k)}\right\rceil \tag{3}$$

称$B(Y,k)$为项集Y的k-支持期望[9]。令$B_{\min}=\min\{B(Y,k)\,|\,q<k\leqslant L\}$，其中$L$是告警交易数据库中交易包含的元素最大值。$q$-项集$Y$的支持度计数满足$\text{Support_Count}(Y)\geqslant B_{\min}(Y)$时被称为加权潜在$q$-项集。

1.2 构造加权频繁模式树

WFP-tree定义如下：

(1) WFP-tree的节点由4个域组成：项集名称、项集支持计数、指向父节点或最左子节点的指针域和指向右兄弟节点或节点链中下一节点的指针域。

(2) 项集的节点链指针定义为：指向树中具有相同项集名称的节点的指针。

(3) WFP-tree根节点结构定义如下：

null	0	null	null

(4)频繁项集的项头表(HeadTable)结构定义如下：

项集名称	节点链指针

从WFP-tree的根节点开始，WFP-tree可以按照下面的算法生成。以告警交易数据库D和最小加权支持度W_{minsup}为算法的输入，输出的是构造好的WFP-tree。算法的执行步骤可归纳如下：

(1) 扫描数据库 D 一次，得到按支持度计数降序排列的加权潜在 1-项集集合 S。

(2) 扫描交易数据库 D，对每一个交易 T_i：

a) 找出 T_i 中的加权潜在 1-项集并按照支持度计数降序排列，得到序列 S_t；

b) 设置指针 temp 始终指向 WFP-tree 中新加入的节点，将序列 S_t 插入 WFP-tree 中。

(3) 以先根次序遍历 WFP-tree 并翻转指向子节点和右兄弟节点的指针，使其分别指向父节点和树中具有相同项集名称的节点并形成节点链。

步骤(2)中定义了一个WFP-tree节点型的动态指针，在构造树的过程中该指针始终指向新加入的节点，这样避免了树的递归生成。另外，WFP-tree是按照从根节点到叶子节点的顺序建立的，但是利用WFP-tree进行规则挖掘的过程则是从叶子节点到根节点的顺序进行的。经过步骤(3)的指针翻转，指向子节点的指针指向了父节点，同时指向右兄弟节点的指针指向了WFP-tree中含有相同节点名称的节点。

1.3 加权关联规则挖掘算法

文献[6]中频繁模式的挖掘需要递归地生成条件模式树，且每产生一个频繁模式就要生成条件模式树。在支持度阈值很小的情况下即使是小数据库也会产生数以万计的频繁模式，动态的生成和释放条件模式树将耗费大量的时间和空间。

本文提出基于WFP-tree的高效关联规则挖掘算法。算法的基本思想是，在挖掘过程中利用WFP-tree挖掘所有的加权频繁项集而不产生条件模式树。在WFP-tree中，将从根节点到项集I所在节点的所有路径称为项集I的前缀路径，那么项集I的所有前缀路径中的项集的集合叫做项集I的条件模式基。WFP-tree包含了数据库中完整的加权频繁模式信息且是对交

易数据库信息的高度压缩，通过节点链可以找到项集α的所有条件模式基。输入WFP-tree、最小加权支持度W_{minsup}、最小置信度minconf和加权潜在1-项集集合S，输出为加权关联规则的集合。算法的执行步骤可归纳如下：

1) 按加权支持计数升序排列 S，对 S 中的每项 n_i：

① 找到 n_i 所有的条件模式基集合 C_{pbset} 和 n_i 所有前缀路径长度的最大值 T_{Max}；

② $C_2 = \{n_i \cup C_1 | C_1 = C_{\text{pbset}}$ 中的加权潜在 1-项集集合$\}$；

③ $(C_2, L_2) = \text{Checking}(C_2)$ // C_k 为加权潜在 k-项集集合；

④ for($k = 3$; $k < T_{\text{Max}}$ && $|C_k|>1$; k++)；

⑤ $C_k = \text{Join}(C_{k-1})$；

⑥ $(C_k, L_k) = \text{Checking}(C_k)$；

⑦ $L = L \cup L_k$ // L_k 为加权频繁 k-项集集合，L 为 L_k 的集合。

2) Rules(minconf, L)。

1.4　算法分析

判断基于WFP-tree的加权关联规则挖掘算法正确与否。

证明：首先证明算法输出的项集是否为加权频繁项集，根据加权频繁项集的定义和加权关联规则挖掘算法中步骤⑥和⑦可知，该算法输出的项集显然是加权频繁项集。然后证明算法是否输出了所有的加权频繁项集。由WFP-tree的构造算法可知WFP-tree是对交易数据库信息的压缩，它包含了交易数库中加权频繁模式的信息。WFP-tree的节点由加权潜在1-项集集合$S=\{\alpha_1,\alpha_2,\cdots,\alpha_k\}$中的各项构成，$S$中元素是按照支持度计数的降序排列的。从WFP-tree中挖掘出来的加权频繁项集应该包括两部分：一是$\alpha_i(i=1,2,\cdots,k)$的加权频繁项集；二是不包含α_i的加权频繁项集。要得到包含α_i的加权频繁项集，只需要在WFP-tree中通过节点链找到α_i的所有前缀路径，构成α_i的条件模式基。然后由加权潜在项集的定义得到α_i的条件模式基中的加权潜在1-项集集合P_1，将α_i与P_1中的元素相连得到加权频繁2-项集的候选项集C_2，并根据加权频繁项集的定义得到加权频繁2-项集。同样道理，对于C_2，找出其中的加权潜在项集集合P_2，连接成C_3，得到加权频繁3-项集。以此类推，可以得到包含α_i的所有加权频繁项集。对于不包含α_i的加权频繁项集，可以处理成包含项集$\alpha_j(j=1,2,\cdots,k, j>i)$的加权频繁项集和不包含项集$\alpha_j$的频繁项集。以此类推，可得到加权频繁项集的全集，并能够产生满足最小置信度的所有关联规则。

所以可以得出基于WFP-tree的加权关联规则挖掘算法理论上是正确的，下面将通过仿真来验证。

2　算法性能测试

本文采用9个网络节点的小型网络(如图1所示)中产生的告警信息为原始告警数据，通过对原始告警数据进行预处理得到能够用于数据挖掘的告警交易数据库。在该网络环境下对MINWAL(O)算法和本文提出的基于频繁模式树的加权关联规则挖掘算法进行性能比较。

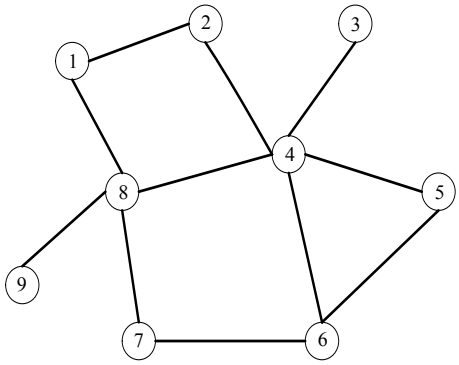

图1　网络拓扑图

(1) 对同一告警交易数据库，比较不同加权支持度下MINWAL(O)算法和本文算法的运行时间。数据库中有500条原始告警信息，经预处理后生成3 974个交易记录，交易最大长度为64，数据库中一共有184个不同的属性。实验结果如图2所示，可以看出本文的算法在时间效率上明显优于MINWAL(O)。当加权支持度为0.15时，本文算法运行时间大约为MINWAL(O)算法的1/7，并且运行时间随着加权支持度减小(加权频繁项集数目增多)而增长的趋势比MINWAL(O)算法平缓的多。

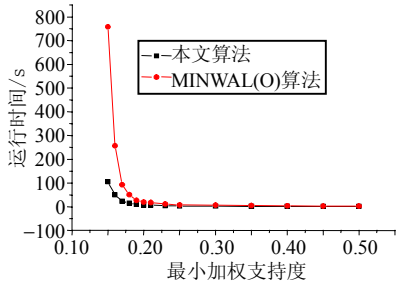

图2　运行时间随加权支持度变化的情况

(2) 算法的可伸缩性比较。可伸缩性指随着数据库的增大，算法的运行时间和存储空间的增长情况。固定最小加权支持度为0.25。算法的时间和空间可伸缩性如图3、图4所示。随着告警数据的增加，两

个算法的运行时间和存储空间都呈增长趋势，但是本文算法的增长速度明显较为缓慢，并且始终优于MINWAL(O)算法。这表明本文算法在运行时间和存储空间上具有更好的可伸缩性。

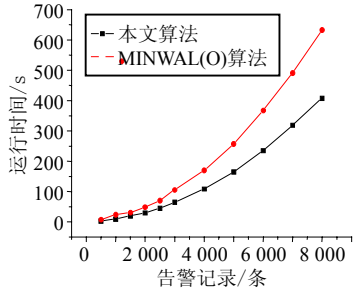

图3 算法时间可伸缩性比较

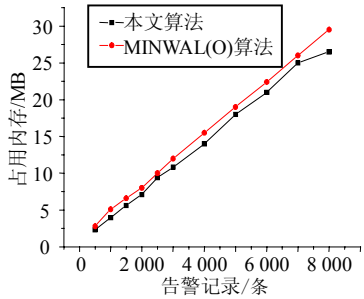

图4 算法空间可伸缩性比较

算法性能测试结果表明，本文提出的算法有以下两个优点：

(1) 算法过程中引入了类似于Apriori方法中的连接、剪枝步骤，不是采用FP-Growth中提出的递归产生条件模式树的方法(条件模式树的数量取决于递归调用的深度,需要大量的附加存储空间)。这样节省了大量的内存资源，提高了算法运行速度和空间效率，能够适应通信网告警信息量大、告警具有突发性这一特定环境下的告警加权关联规则挖掘。

(2) 在进行加权关联规则挖掘的过程中以WFP-tree为基础而不需要重复遍历告警交易数据库，大大提高了算法的时间效率。

3 结 论

针对通信网告警信息具有不同权重、告警信息量大和告警具有突发性等特点，提出了一种高效的基于加权频繁模式树的通信网告警加权关联规则挖掘算法。算法分为两个部分：构造WFP-tree和基于WFP-tree的加权关联规则挖掘。算法性能测试表明，与MINWAL(O)算法相比较，本文算法的挖掘效率有明显的提高，并且具有更好的时间和空间可伸缩性，对通信网告警相关性分析，快速、准确的定位告警故障根源具有很大的意义。

参 考 文 献

[1] GARDNER R D, HARLE D A. Pattern discovery and specification techniques for alarm correlation[C]//Proc of the IEEE NOMS'98 Conference. New Orleans, LA, USA: IEEE Press, 1998: 713-722.

[2] WU Y Y, DU S G, LUO W, Mining alarm database of telecommunication network for alarm association rules[C]// Proc of 11th Pacific Rim International Symposium on Dependable Computing. Washington, DC, USA: IEEE Computer Society, 2005: 281-286.

[3] AGRAWAL R, SRIKANT R. Fast algorithm for mining association rules[C]//Proc of the 20th VLDB Conference, Santiago. Chile: Morgan Kaufmann/ACM, 1994: 487-499.

[4] SARAWAGI S, THOMAS S, AGRAWAL R. Integrating association rule mining with relational database system: Alternatives and implications[C]//Proc of ACM SIGMOD International Conf on Management of Data. Seattle, Washington: ACM Press, 1998: 343-354.

[5] LI Z C, HE P L, LEI Ming. A high efficient AprioriTid algorithm for mining association rule[C]//Proc of 4th International Conference on Machine Learning and Cybernetics. Guangzhou: IEEE Press, 2005: 18-21.

[6] Han J W, Pei J, Yin Y W, et al. Mining frequent patterns without candidate generation: a Frequent-Pattern tree approach[J]. Data Mining and Knowledge Discovery, 2004, 8(1): 53-87.

[7] WANG W, YANG J, YU P. Efficient mining of weighted association rules (WAR)[C]//Proc of the ACM SIGKDD Conf on Knowledge Discovery and Data Mining. Washington DC, USA: ACM Press, 2000: 270-274.

[8] OUYANG Wei-min, ZHENG Cheng, CAI Qing-sheng. Discovery of weighted association rules in databases[J]. Journal of software, 2002,12(4): 612-619.

[9] CAI C H, FU A W, CHENG C H, et al. Mining association rules with weighted items[C]//Proc of IEEE International Database Engineering and Applications Symposium. Cardiff: IEEE Computer Society, 1998: 68-77.

[10] 肖海林, 李兴明. 层次分析法在通信告警加权关联规则挖掘中的应用研究[J]. 电信科学, 2006, 22(11): 36-39.
XIAO Hai-lin, LI Xing-ming. Study of analytic hierarchy process applied to the alarm correlation analysis in communication networks[J]. Telecommunication Science, 2006, 22(11): 36-39.

编 辑 张 俊