



基于综合相关置信度的告警关联规则挖掘^{*}

高会生,李英敏

(华北电力大学电子与通信工程系 保定 071000)

摘要:采用置信度融合算法计算相邻告警的综合相关置信度,并根据计算结果形成告警关联事务,增强告警事务内告警之间的相关性,降低对支持度的选择要求,有效挖掘出“低支持度-置信度”告警关联规则。利用h-置信度理论筛选有意义的告警关联规则,并通过实验比较了传统方法和本文方法的执行效率和挖掘正确率。

关键词:告警关联规则;综合相关置信度;h-置信度

doi: 10.3969/j.issn.1000-0801.2014.10.017

Mining Alarm Association Rules Based on Comprehensive Related Confidence

Gao Huisheng, Li Yingmin

(Department of Electronic and Communication Engineering of North China Electric Power University, Baoding 071000, China)

Abstract: The traditional association rules mining algorithms preferred to applying single sliding time window, and the mined alarm association rules were high support-high confidence. To solve this problem, a method was proposed to calculate comprehensive related confidence between the time adjacent items by the confidence fusion algorithm. Then generated alarm transactions were generated based on the calculated results, which enhanced alarm correlation within a transaction. The proposed approach can reduce the choice of support requirement and mine low support-high confidence association rules effectively. Based on h-confidence theory, interesting alarm association rules were screened. Experimental results showed the feasibility and effectiveness of this approach.

Key words: alarm association rule, comprehensive related confidence, h-confidence theory

1 引言

WINEPI算法^[1]广泛应用于告警关联规则挖掘,其基本思想是利用滑动时间窗获得告警关联事务库,并根据“高支持度—高置信度”获得强告警关联规则。WINEPI算法的关键问题是设定滑动时间窗口大小及选择合适的支持度和置信度阈值。本文通过对某电力通信专网的历史告警库统计发现,告警的持续时间以及出现时间随机性较强,如图1所示。显然,这种告警持续时间和告警发生时刻的随

机性增加了滑动时间窗口的确定难度。对全网采用统一的告警时间窗口可能存在如下问题:若窗口选取过小,会丢失相关告警信息;若选取过大,不仅会引入虚假规则,而且由于引入了无关数据,降低了算法效率^[2]。针对此问题,参考文献[2]利用告警窗口宽度对告警数据库进行时间跨度的约束,根据实际告警数据动态调整窗口大小,但是该算法需要统计所有告警类型的持续时间分布情况,不适用于低频率告警,且较复杂。参考文献[3]提出了Spatio-Temporal Apriori算法,实现了关联规则的时空分析,

^{*} 中央高校基本科研业务费专项资金资助项目(No.13XS32)

但是该方法仅简单地将某一时间段内空间位置相同的项进行关联,忽略了时空关系的定量影响,关联粒度较粗。而其他研究^[4-6]均直接借鉴 WINEPI 算法用于告警关联的前期阶段,并没有针对此问题做出建设性改进。

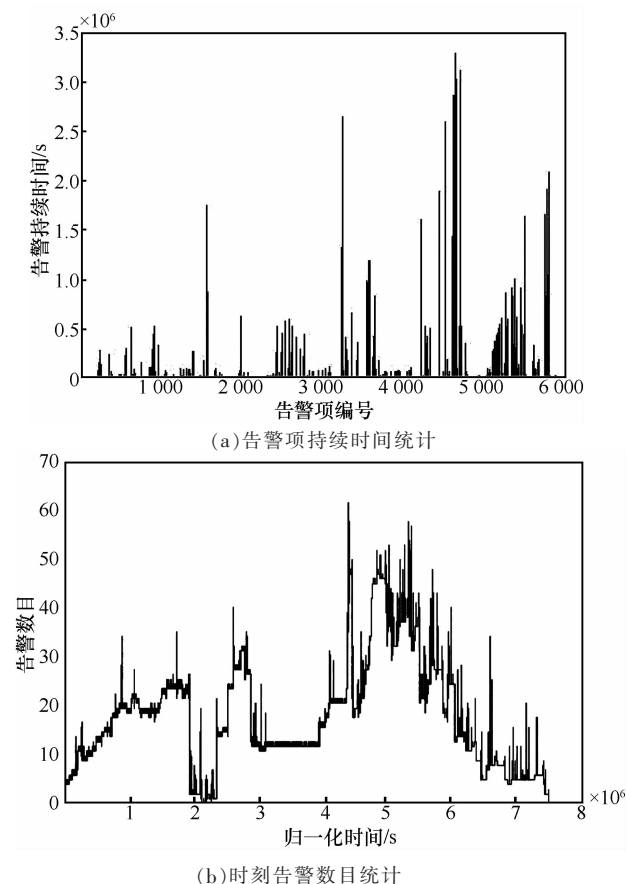


图1 告警时间分布规律统计

对此,本文提出了基于综合相关置信度——h-置信度的告警关联规则挖掘算法。算法利用告警的时间属性和网络拓扑关系实现多证据置信度融合,计算告警历史库中相邻告警项的综合相关置信度,进而形成告警关联事务库,实现了对告警事务的时空细粒度约束。这样可以使事务内告警之间存在较高的相关性,降低了对时间窗口的选取要求。在这种情况下,即使支持度阈值较小也不会引入过多的虚假告警关联规则,可以有效发现低支持度-高置信度的告警关联规则。本文还结合 FP-Growth 算法^[7]和 h-置信度理论^[2],快速筛选出有意义的告警关联规则,实现了高效挖掘。

2 告警关联算法描述

两条告警之间的关联关系可以用五元组 $AR=(A_i, A_j,$

$TI, NT, R)$ 表示。其中, i 和 j 均为正整数, A_i 和 A_j 分别表示两条在时间上相邻的告警项; TI 为两条告警项的时间间隔,用绝对值表示; NT 是对产生两条告警源站的网络拓扑关系的量化值。在这里,若源站相同,则 $NT=1$;若源站不同,但相邻,则 $NT=2$;若源站不满足上述两个条件,则 $NT=3$ 。在五元组中, R 代表两个告警项的综合相关置信度值。

本文中,告警关联规则挖掘算法主要包含以下 3 个步骤。

(1) 计算相邻告警的综合相关置信度,并根据综合相关置信度形成告警关联事务库。

(2) 挖掘频繁告警序列。

(3) 筛选告警关联规则。在进行规则挖掘之前,算法需要的输入数据包括:RD(raw database)、h-置信度阈值(min_C)、支持计数阈值(min_S)、相关时间阈值(TR_1 和 TR_2)以及 NTM(network topology matrix,网络拓扑关系矩阵),规则挖掘算法运行结束后产生 ATD(alarm transaction database,告警事务库)和 ACRB(alarm correlation rule base,告警关联规则库)。下面对告警关联规则挖掘算法的主要步骤进行说明。

2.1 计算告警综合相关置信度

2.1.1 综合相关置信度计算公式

广义上讲,置信(confidence)指的是正确的概率,置信度(confidence measure)是评价这个概率的一种量度,表示某一事件的可靠程度^[8]。在这里,告警的综合相关置信度是指在时空双重条件约束下两条告警相关的可靠程度。利用综合相关置信度对告警的相关性进行度量,可以有效减小由单一证据获得相关置信度的不确定性。参考文献[9]给出了置信度综合公式。

假设存在多条证据共同作用于一个信息 H,例如 IF E1 THEN H 和 IF E2 THEN H,其置信度分别为 CF_1 和 CF_2 ,则信息 H 的置信度综合公式如下。

$$CF_{\text{comb}}(CF_1, CF_2) = \begin{cases} CF_1 + CF_2(1 - CF_1), & CF_1 > 0, CF_2 > 0 \\ \frac{CF_1 + CF_2}{1 - \min(|CF_1|, |CF_2|)}, & CF_1 > 0, CF_2 < 0 \text{ 或 } CF_1 < 0, CF_2 > 0 \\ CF_1 + CF_2(1 + CF_1), & CF_1 < 0, CF_2 < 0 \end{cases} \quad (1)$$

其中,当 CF_1 或 CF_2 大于零时,代表 E1 或 E2 对 H 起支持作用,为有利证据;否则起反对作用,为不利证据。

本文结合某电力通信专网告警的特点,定义告警综合相关置信度公式如下。



$$R(A_i, A_j) = CF(CF_1, CF_2)$$

$$= \begin{cases} CF_1, & CF_1=1 \text{ 或 } CF_2=0 \\ CF_1+CF_2(1-CF_1), & 0 < CF_1 < 1, CF_2 > 0 \\ \frac{CF_1+CF_2}{1-\min(|CF_1|, |CF_2|)}, & 0 < CF_1 < 1, CF_2 < 0 \\ CF_2, & 0 < CF_1 < 1, CF_2=0 \end{cases} \quad (2)$$

CF_1 是由告警时间间隔 TI 确定的告警相关置信度, 如式(3)所示。

$$CF_1 = \begin{cases} 1, & 0 \leq TI \leq TR_1 \\ \frac{TR_1}{TI}, & TR_1 \leq TI < TR_2 \\ 0, & TI \geq TR_2 \end{cases} \quad (3)$$

只考虑告警时间对相关性的影响, 当 $TI < TR_1$ 时, $CF_1=1$, 代表仅根据 TI 可以确定两个告警是相关的; 当告警时间间隔 TI 介于 TR_1 和 TR_2 之间时, $CF_1=TR_1/TI$, 此时将告警的相关置信度模糊化, 告警的相关性由 CF_1 、 CF_2 共同决定; 当 $TI > TR_2$ 时, $CF_1=0$, 认为两个告警不相关。

CF_2 根据 CF_1 的值及告警网络拓扑关系确定, 如式(4)所示。

$$CF_2 = \begin{cases} -0.5, & 0 < CF_1 < 1, NT=1 \\ 0.5, & 0 < CF_1 < 1, NT=2 \\ 1, & 0 < CF_1 < 1, NT=3 \end{cases} \quad (4)$$

根据统计得, 由同一个站产生的告警平均时间间隔较小, 在一定的时间范围内, 随着 TI 变大, 两个告警相关的概率减小, 此时 CF_2 相对于 CF_1 为不利证据, 取值为负。而由相邻源站产生的告警平均时间间隔较大, 在一定的时间范围内, 随着 TI 增大, 两个告警相关的概率变大, 则此时 CF_2 相对于 CF_1 为有利证据, 取值为正。根据统计学理论, 在有其他背景知识的支持下取 $|CF_2|=0.5$ 。如果产生告警的两个源站既不相同也不相邻, 有 $CF_2=0$, 即一般情况下, 告警不能越站引发, 在没有较小时时间间隔的保证下, 两个告警相关的概率非常小, 可以忽略不计。

TR_1 、 TR_2 按照置信水平经验值获得。对全网的相邻告警时间间隔进行统计, 剔除较大的时间间隔 (认为相邻两告警项的时间间隔过大时, 两告警项一定不相关), 最多有 90% 的告警时间间隔小于或等于 TR_1 , 最多有 95% 的告警时间间隔小于或等于 TR_2 。

2.1.2 综合相关置信度计算步骤

(1) 初始化 $i=1, j=i+1$ 。

(2) 判断两条告警 A_i, A_j 的 TI , 若 TI 大于 TR_2 , 则赋值 $R=0$; 若小于 TR_1 , 赋值 $R=1$; 否则, 执行步骤(3)。

(3) 判断 NT 的值, 如产生两告警的通信站既不相同也不相邻, 则赋值 $R=0$; 否则, 执行步骤(4)。

(4) 根据式(4)计算 R 值, 若 $R \geq 0.9$, 则赋值 $R=1$; 否则, 赋值 $R=0$ 。

(5) 判断 R 值, 若 $R=1$, 则将两个告警放入一个告警事务, 赋值 $i=i+1, j=j+1$, 执行步骤(7); 否则, 执行步骤(6)。

(6) 产生一个新事务, 将 A_i 作为新事务的第一项, 赋值 $i=i+1, j=j+1$, 执行步骤(7)。

(7) 判断是否到达历史告警库的末尾。若是, 则结束; 否则, 执行步骤(2)。

2.1.3 算法复杂度

每增加一条告警项, 基本操作多执行一次, 循环嵌套次数没有增加, 所以算法的时间复杂度 $T(n)$ 随着历史告警库规模 n 的增大呈线性阶 $O(n)$ 特点。

2.2 筛选告警关联规则

FAS(frequent alarm sequence, 频繁告警序列)支持计数大于或等于 $\min_S$ 的告警序列, 本文利用 FP-Growth 算法对 ATD 进行挖掘。产生 FAS 的告警项都是按照支持度由大到小 $\min_S$ 的顺序排列。根据这一特点, 本文采用 h-置信度, 快速筛选出 IACR (interesting alarm correlation rule, 有意义的告警关联规则), IACR 指支持计数和 h-置信度都满足阈值的 FAS。若只满足 $\min_S$, 不满足 $\min_C$, 为 FACR (false alarm correlation rule, 虚假告警关联规则)。

定义 1 (告警项支持计数) 告警项 A 的支持计数指 ATD 包含告警项 A 的数目, 记为 $N(A)$, 告警序列 $X_n=(A_i, A_j, \dots, A_n)$ 的支持计数指 ATD 包含 X_n 的数目, 记为 $N(X_n)$, 其中, i, j, n 均为正整数。

定义 2 (h-置信度) 频繁告警序列 X_n 的 h-置信度的定义如式(5)所示。

$$C(X_n) = \frac{N(X_n)}{\max_{1 \leq i \leq n} (N(A_i))} \quad (5)$$

根据参考文献[2]证明可知, h-置信度具有反单调性, 即有 $C(X_{n+1}) \leq C(X_n)$ 。

根据这一性质可得, 当 $C(X_n) \geq \min_C$ 且 $C(X_{n+1}) < \min_C$ 时, X_n 的任意子集 X_k 均有 $C(X_k) \geq \min_C$ 成立, 其中 $2 \leq k \leq n$ 且为正整数, 此时有 X_n 及其子集为有意义的告警关联规则; X_{n+1} 及其超集的 h-置信度均小于 $\min_C$, X_{n+1} 及其超集为虚假告警关联规则。

3 实验结果

为了对所提方法的有效性和可行性进行验证,本文选择电力通信专网的告警库作为实例,开展告警关联规则数据挖掘。实际通信网络是基于 SDH 技术的光传输网,网络总共有 212 个通信站,每一个站都安装一台 SDH 设备。本文采集了 19 992 条告警形成源告警数据库。在 i3 3.30 GHz CPU、4 GB 内存、Windows 8.0、MATLAB 2013.a 环境下获得告警关联事务库,在 Windows XP VC++环境下采用 FP-Growth 算法对生成的告警关联事务库进行频繁模式挖掘。将获得的关联规则存储在 Windows 8.0 Microsoft Office Excel 2007 中。

3.1 执行时间对比

比较由本方法和基于滑动时间窗的方法(以下简称传统方法)在获取告警关联事务库时所需要的时间长短。由于所用 MATLAB 环境为解释性语言,同一程序的执行时间具有不稳定性,本文采用 3 次实验的平均值作为最终结果进行比较。本文方法的 TR_1 设置为 180 s、300 s、500 s、660 s 和 780 s,相应的 TR_2 为 360 s、590 s、978 s、1 320 s 和 1 560 s,传统方法的时间窗口 TW 也分设 180 s、300 s、500 s、660 s 和 780 s。结果如图 2 所示。

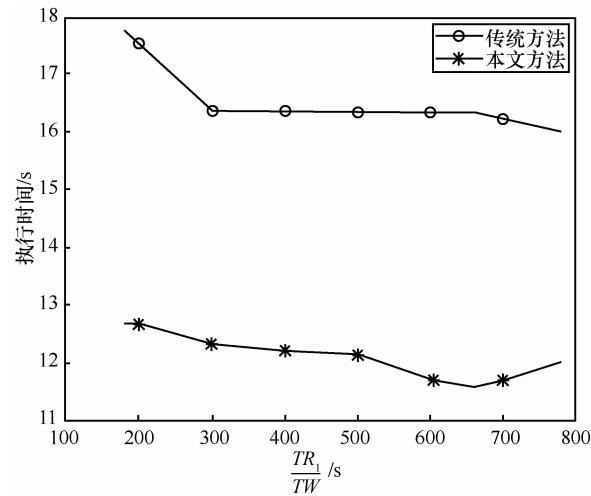


图 2 执行时间对比

由于历史告警库的告警数目庞大,在形成告警事务的过程中,程序的迭代次数决定了程序的执行效率。由图 2 可得,当 TW 变大时,由传统方法形成的告警事务数减少,程序的执行时间相应降低;随着 TR_1 的增大,本文方法的执行时间比传统方法变化平稳,且相对较短。这是因为传统方法的相邻窗口之间部分重叠,在一定的时间窗口范围

内形成的告警事务数目比实际数目多得多。由本文方法形成告警事务时,考虑到“网络拓扑结构”对告警相关性的影响,使告警事务内告警项之间的差异性更小,事务间的告警之间差异性更大,这样形成的告警事务数目更加符合实际情况,且比传统方法少得多,进而程序的迭代次数相对减少,执行时间降低。由不同 TR_1/TW 对同一告警历史库进行挖掘形成的告警事务数目见表 1。由表 1 可得,在 TR_1/TW 值相同的情况下,由传统方法获得告警事务数目明显多于本文方法,且数目随着时间窗口的增大而减少,从而印证了前文对执行时间变化原因的分析。

表 1 告警事务数目对比

$\frac{TR_1}{TW}/s$	传统方法获得的告警事务数目	本文方法获得告警事务数目
780	1 603	1 221
660	1 605	1 156
500	1 691	1 289
300	1 864	1 445
180	2 079	1 507

3.2 不同 TR_1 、 TR_2 和 TW 下的比较

在不同 TR_1 、 TR_2 和 TW 下,给出本文方法和传统方法获得的 FAS、FACR、IACR 数目变化情况。

其中,最小支持计数设为 20, h -置信度阈值设为 50%。

TR_1 、 TR_2 和 TW 的取值和第 3.1 节中相同,比较结果如图 3 所示。

分析图 3(a)可得,由传统方法获得的 FAS 的数目多于本文方法,且随着 TW 变大,FAS 数目变化幅度较大;而由本文方法得到的 FAS 数目变化稳定,但数目较少。这是因为本文方法在进行告警综合相关置信度计算时,丢失了一定的信息,而传统方法在进行时间窗口重叠滑动时增加了告警事务的信息量,导致两种方法获得的 FAS 数目不同。传统方法的 TW 是决定告警事务内容的唯一因素,所以当 TW 变化时,FAS 数目变化幅度较大,且呈单调递增形势增长,即 TW 越大,引入的信息越多;本文方法实现了告警事务时空双重约束,FAS 的数目变化跟 TR_1 、 TR_2 及 NT 3 个因素有关,所以 FAS 数目变化随着 TR_1 单一因素的改变趋势不明显,而且不具有单调性。

分析图 3(b)、图 3(c)得,由两种方法获得的 IACR 数目相近,而传统方法的 FACR 数目明显多于本文方法的 FACR。说明本文方法的挖掘准确率要高于传统方法。这是因为通过告警综合相关置信度的计算,使事务内的告警相

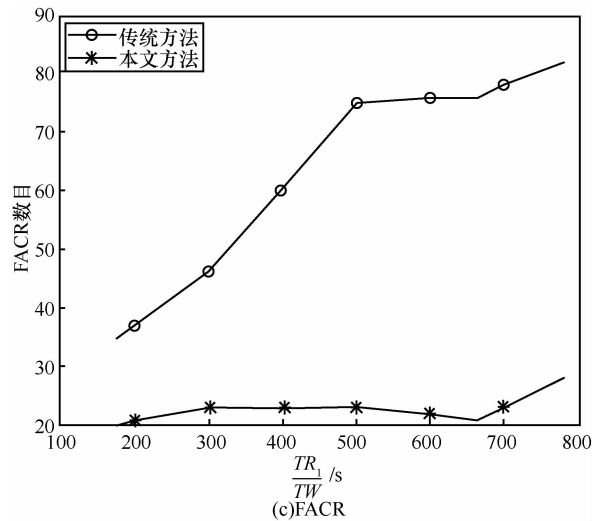
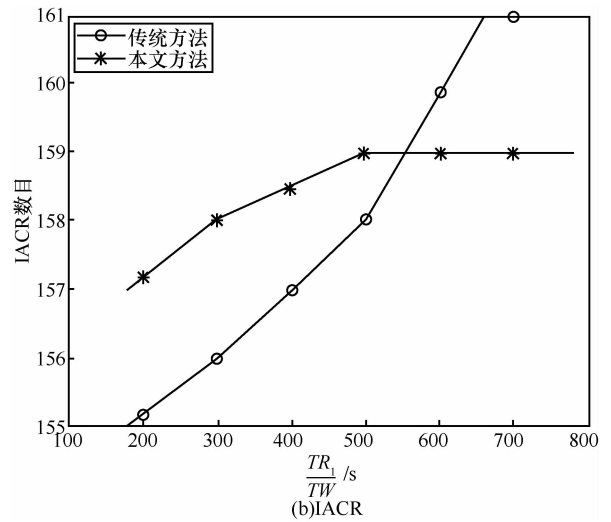
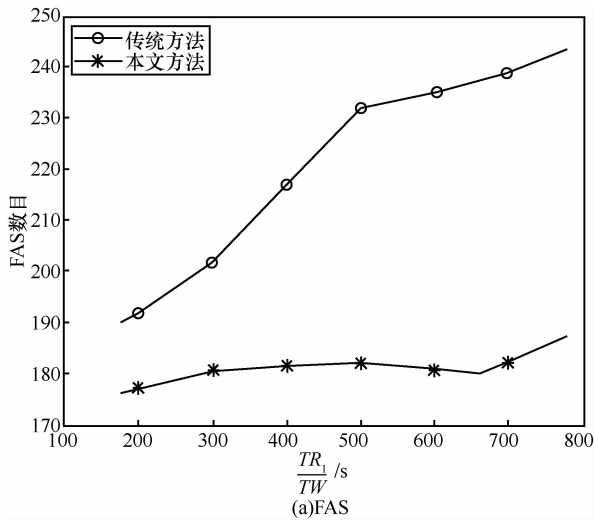


图3 FAS数目、IACR数目、FACR数目比较

关性更强,事务间的告警相关性更弱,即使告警事务的内容更加符合客观事实,提高了挖掘结果的准确性。而传统方法的结果直接由滑动时间窗口决定,当滑动窗口过小时,会丢失信息;过大时,引入的无关信息增多,时间窗口的选择不易把握。综上所述,本文方法形成告警关联事务库的效率和挖掘准确率都高于传统方法,证明了本文方法的可行性和有效性。

3.3 挖掘结果的信息完备性比较

设 $S_1=\{t_{1i}|i \text{ 为正整数}, t_{1i} \text{ 为由传统方法获得的 IACR}\}$, $S_2=\{t_{2i}|i \text{ 为正整数}, t_{2i} \text{ 为由传统方法获得的 IACR}\}$, $S=\{t_i|i \text{ 为正整数}, t_i \in S_1, \notin S_2\}$ 。 $CICR$ (correct information coverage rate) 为有效信息覆盖率,则有 $CICR_{12}=|S|/|S_1|$ 为传统方法相对于本文方法的 $CICR$, 此时称本文方法为参考方法; $CICR_{21}=|S|/|S_2|$ 为本文方法相对于传统方法的 $CICR$, 此时称

传统方法为参考方法。 $S_{21}=\{t_{2i}|i \text{ 为正整数}, t_{2i} \in S_1, t_{2i} \notin S_2\}$, $S_{12}=\{t_{1i}|i \text{ 为正整数}, t_{1i} \in S_2, t_{1i} \notin S_1\}$ 。 $CILR$ (correct information loss rate) 为有效信息丢失率,则有 $CILR_{21}=|S_{21}|/|S_2|$ 为本文方法相对于传统方法的 $CILR$, 此时称传统方法为参考方法; $CILR_{12}=|S_{12}|/|S_1|$ 为传统方法相对于本文方法的 $CILR$, 此时称本文方法为参考方法。选择与第 3.1 节相同参数, 获得结果见表 2。

有效信息覆盖率越高, 有效信息丢失率越低, 说明所用方法的挖掘结果相对于参考方法越完善。由表 2 得知, 本文方法相对传统方法有效信息覆盖率达到 98% 以上, 有效信息丢失率保持在 1.89% 以内, 两个结果均满足实际工程要求。同时发现, 在不同的 TR_1/TW 取值下有 $CILR_{12}>0$, $CILR_{21}\geq 0$ 成立, 说明, 通常情况下, 两种方法的挖掘结果不能完全相互覆盖 (当 $CILR_{12}=0$ 或 $CILR_{21}=0$ 时称完全相互覆

表 2 有效信息覆盖率和有效信息丢失率对比

$\frac{TR_1}{TW}/s$	$ S_1 $	$ S_2 $	$ S $	$CICR_{12}$	$CICR_{21}$	$ S_{21} $	$ S_{12} $	$CILR_{21}$	$CILR_{12}$
780	161	159	158	98.13%	99.37%	3.	1	1.89%	0.62%
660	161	159	158	98.13%	99.37%	3	1	1.89%	0.62%
500	158	159	156	98.73%	98.11%	2	3	1.26%	1.90%
300	156	158	156	100%	98.73%	0	2	0.00%	1.28%
180	155	157	153	98.71%	97.45%	2	4	1.27%	2.58%

盖),考虑到挖掘结果的完备性,本文将把 $S \cup S_{12} \cup S_{21}$ 作为最终挖掘结果呈现给网管人员做进一步专家分析。

3.4 挖掘结果的实际物理意义

综合第 3.1 节和第 3.2 节的结果,本文选择 $TR_1=660\text{ s}$, $TR_2=1\ 320\text{ s}$, $\min_S=20$, $\min_C=50\%$, $TW=660\text{ s}$ 作为挖掘参数,获得的告警关联规则见表 3。

表 3 告警关联规则

告警关联规则序列	h-置信度
CP_OSPF_NB_DOWN.CP_TEL_DOWN	75.80%
...	...
CP_OSPF_NB_DOWN.CP_TEL_DOWN. CP_OSPF_CL_DOWN.CP_CC_DOWN. CP_SER_NOT_OR.CP_TEL_DEG.CP_SER_INT	50.12%

表 3 中每一行代表一条告警关联规则,如第一行表示告警关联序列在 CP_TEL_DOWN 或 CP_OSPF_NB_DOWN 出现的情况下,CP_TEL_DOWN、CP_OSPF_NB_DOWN 同时出现的最小概率为 75.80%,即在 OSPF 链路中断的情况下发生 TE 链路中断的概率为 75.80%,且 CP_OSPF_NB_DOWN 出现的概率大于 CP_OSPF_NB_DOWN。

4 结束语

本文首次引入置信度融合算法对源告警数据库中的告警项进行综合相关置信度的计算,并根据计算结果形成告警事务库,首次将告警的拓扑关系分为有利证据和不利证据融入告警综合相关置信度,使计算结果更加符合客观事实,降低了对相关时间及支持度的要求,提高了挖掘结果的正确率。同时,根据 FP-Growth 算法挖掘出的频繁告警序列的告警项按照支持度由大到小排列的特点,利用 h-置信度理论的反单调性快速筛选出有意义的告警关联规则,去除虚假告警关联规则,实现了高效的告警关联规则挖掘算法。

参考文献

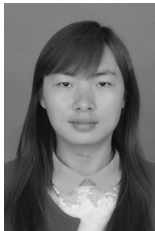
1 Mannila H, Toivonen H, Verkamo I A. Discovery of frequent

episodes in event sequences. Data Mining and Knowledge Discovery, 1997(1)
2 徐前方,肖波,郭军. 挖掘电信告警关联模式方法. 北京邮电大学学报, 2011, 34(2):86~89
3 夏英,张俊,王国胤. 时空关联规则挖掘算法及其在 ITS 中的应用. 计算机科学, 2011, 38(9):173~176
4 Lagzian S, Amiri F, Enayati A. Frequent item set mining-based alert correlation for extracting multi-stage attack scenarios. Proceedings of the 6th International Symposium on Digital Object Identifier, Iran, Tehran, 2012
5 田志宏,张永铮,张伟哲等. 基于模式挖掘和聚类分析的自适应告警关联. 计算机研究与发展, 2009, 46(8): 1304~1315
6 吴简,李兴明. 基于关联规则的分布式通信网告警相关性研究. 计算机科学, 2009,36(11): 204~207
7 Han J W. Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, New York, USA, 2000
8 刘镜,刘加. 置信度的原理及其在语音识别中的应用. 计算机研究与发展, 2000, 37(7):882~889
9 王海栋,郑晓庆,张红俊. 基于置信度理论的网络知识融合系统和应用. 计算机系统应用, 2011, 20(1): 1~6

[作者简介]



高会生,男,博士,华北电力大学电子与通信工程系教授、系主任,主要研究方向为电力系统通信。



李英敏,女,华北电力大学电子与通信工程系硕士生,主要研究方向为电力通信专网告警关联。

(收稿日期:2014-08-13)