

基于双约束滑动时间窗口的告警预处理方法研究

李彤岩¹, 李兴明²

(1. 成都信息工程学院 通信工程系, 成都 610225; 2. 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: 在通信网告警相关性分析中, 针对传统时间窗口提取告警数据效率低的问题, 提出了一种基于双约束滑动时间窗口的告警预处理方法。在双约束的条件下, 滑动时间窗的窗口宽度和滑动步长能够根据告警序列的实际情况自动地调整, 并选取时间段的中点作为参照点。实验证明, 采用双约束滑动时间窗划分算法可以获得比传统方法更高的数据提取效率, 并且能够有效去除噪声数据, 非常适合于通信网告警数据的预处理。

关键词: 时间窗口; 滑动步长; 双约束; 告警相关性分析; 告警事务

中图分类号: TP393 **文献标志码:** A **文章编号:** 1001-3695(2013)02-0582-03

doi:10.3969/j.issn.1001-3695.2013.02.073

Study of alarm pretreatment based on double constraint sliding time window

LI Tong-yan¹, LI Xing-ming²

(1. Dept. of Communication Engineering, Chengdu University of Information Technology, Chengdu 610225, China; 2. Key Laboratory of Broadband Optical Fiber Transmission & Communication Networks of Ministry of Education, University of Electronic Science & Technology of China, Chengdu 610054, China)

Abstract: In communication alarm correlation analysis, traditional time window has low efficiency on extracting data. This paper first proposed a novel double constraint based sliding window method. With double constraint condition, the window width and sliding step of the time window could automatically adjust according to the actual situation of alarm sequences, and the mid-point of the time period was selected as a reference point. The performance testing of the algorithm indicates that this method has higher extract efficiency than traditional methods, which is suitable for alarm data in telecommunication networks.

Key words: time window; sliding step; double constraint; alarm correlation analysis; alarm transactions

数据挖掘技术是一种适合于通信网告警相关性分析^[1-3]的方法。然而, 通信网的原始告警数据往往存在着信息冗余、数据不完整、时间不同步等问题, 直接对原始告警进行挖掘不仅效率低, 而且准确性较差。为了提高挖掘的效率和准确性, 就必须对原始告警数据进行转换, 以生成适合于挖掘的告警事务数据。由于原始告警一般具有时间不同步性, 所以历史告警记录中的时间一般存在一定的偏差, 滑动时间窗方法^[4-7]正好解决了这一问题。设定滑动时间窗时, 认为在同一窗口内发生的告警事件都是同时发生的。设定滑动时间窗的作用有: 可以将告警数据库转换为告警事务数据库, 同一窗口内的告警被当成是在同一个告警事务内; 有些告警会在短暂的时间内重复出现多次, 为了消除冗余的告警, 在同一时间窗内出现的相同类型的告警只记录一次; 滑动时间窗可以提取出告警的相关性信息, 既充分提取了发生频率频繁的告警, 又不会遗漏掉告警之间的关联性。例如两个相邻发生的告警, 即使在前一个时间窗没有被划分在同一告警事务内, 但随着时间窗的滑动, 仍然有可能被划分到下一个告警事务内。由此可见, 窗口宽度直接影响了每个事务内的告警数, 而滑动步长则关系到告警事务库中的事务数。所以设定一个大小合适的窗口和滑动步长对告警关联规则的挖掘有着十分重要的作用和意义。

已有的时间窗研究大多是设定固定的窗口宽度^[7,8]。这

种方法比较简单、容易操作。但是告警序列的特点是具有突发性, 可能在一段时间内连续出现的告警频率非常频繁, 而一段时间过后告警出现得就不那么频繁了。在这种情况下, 如果设定固定宽度的时间窗, 就会造成告警频繁出现的时候不能够充分地捕捉到告警事务、而告警不频繁出现的时候滑动时间窗的操作效率又很低的结果。

本文设计了一种基于双约束的滑动时间窗来提取告警事务, 其中的双约束是指对滑动时间窗所在时间段划分的约束, 即窗口所在的时间段的段内差异最小和段间差异最大。在双约束的条件下, 滑动时间窗的窗口宽度和滑动步长会根据告警序列的实际情况自动地调整, 以更高的效率提取告警事务。

1 相关概念

1.1 滑动时间窗定义

滑动时间窗是处理时间序列常用的方法, 首先介绍一下滑动时间窗的相关概念。

定义1 时间窗。给定一个告警事件集 E , 一个告警序列 $S = \{s, T_s, T_e\}$ 。其中, T_s 是告警事件序列开始的时间, T_e 是告警事件序列结束的时间, s 是一个告警的递增序列, 即 $s = (e_1, t_1), (e_2, t_2), \dots, (e_n, t_n)$, 其中 $e_i \in E, T_s \leq t_i \leq T_e, i = 1, \dots, n$ 。

收稿日期: 2012-05-21; 修回日期: 2012-06-28

作者简介: 李彤岩(1980-), 女, 河南洛阳人, 讲师, 博士, 主要研究方向为通信网管理和数据通信(sunny60138800@yahoo.com.cn); 李兴明(1954-), 男, 云南人, 博导, 教授, 主要研究方向为通信网络管理。

子序列 $S_w = \{w, t_s, t_e\}$ 称为告警事件序列 S 的一个时间窗口, 其中 $t_s > T_s, t_e < T_e, w = \{w \subseteq S | t_s < t < t_e\}$ 。 $t_e - t_s$ 称为窗口宽度, 记为 W 。

定义2 窗口滑动步长 s 。假设告警事件序列 (简称告警序列) 在当前窗口内起始时间为 T_i , 结束时间为 T_e , 则 $T_e - T_i = W$ 。窗口滑动一个步长 s 后, 新窗口内起始和结束时间分别为 $T_i + s$ 和 $T_e + s$ 。

定义3 告警情景发生的最小间隔 Q 。给定一个告警序列 S 和序列内的一个情景 M (情景由两个或两个以上告警事件组成), 如果同时满足条件: a) 情景 M 在间隔为 Q 的窗口出现内发生; b) 在间隔 Q 的任意子间隔内, 情景 M 都不发生, 则称时间间隔 Q 为序列内情景 M 发生的最小间隔。

举例说明定义3, 给出一组告警递增序列: $(A, t_1), (A, t_2), (B, t_3), (B, t_4)$, 其中 $t_1 < t_2 < t_3 < t_4$ 。对于情景 AB 来说, $t_3 - t_2$ 为最小间隔, 而 $t_3 - t_1$ 和 $t_4 - t_2$ 都不是最小间隔。

定义4 时间窗内告警情景发生的最大间隔 P 。在一个时间窗内, 对于情景 M 来说, 间隔最大的时间段称为最大间隔。

针对于定义3中的例子, 对于情景 AB 来说, $t_4 - t_1$ 是最大间隔, 而 $t_3 - t_1, t_3 - t_2$ 和 $t_4 - t_2$ 都不是最大间隔。

1.2 告警序列时间段划分的双约束条件

告警序列是一组不均匀分布的离散数据, 但是相关性比较大的告警数据总是比较密集地出现在一个时间段内。如果将一个相对密集的时间段查找出来, 然后分别设定各自的时间窗宽度和滑动步长, 就可以提高提取告警事务的效率和有效性, 使得提取出来的告警事务之间的相关性程度比较高, 进而提高关联规则挖掘的效率。

对于给定的一个告警序列 $s = (e_1, t_1), (e_2, t_2), \dots, (e_n, t_n)$, 其中含有 n 个时间点, 将这个告警序列构造成为 k 个时间段划分, $k \leq n$, 则这 k 个时间段划分须满足下面的条件: a) 每一个时间段内至少包含有一个告警事件; b) 每一个告警事件都属于且仅属于这一个时间段。对于给定的 k , 首先给出一个初始的划分, 以后通过反复迭代的方法更新划分, 使得每一次更新后的划分方案都比前一次好, 直到最小化所有对象与参照点之间的相异度之和为止。

在告警序列的时间段划分中, 所谓划分方案好的标准是: 同一时间段内告警事件距离越近越好, 而不同时间段内的告警事件距离越远越好。距离远近或者相异度都是划分时间段的评价函数。时间段划分的约束函数定义有两个原则: 每个时间段内应该是紧凑的, 各个时间段之间距离应该尽可能远。分别用段内差异 $w(t)$ 和段间差异 $L(t)$ 来表示时间段的紧凑性和距离。

段内差异函数定义的最简单方法是计算时间段内每一个时间点到它所属的时间段中心点的距离的平方和:

$$w(t) = \sum_{i=1}^k w(t_i) = \sum_{i=1}^k \sum_{t \in t_i} d(t, t_i)^2 \quad (1)$$

段间差异函数定义为各个时间段中心点之间的距离:

$$L(t) = \sum_{1 \leq j < i \leq k} d(\bar{t}_i, \bar{t}_j)^2 \quad (2)$$

时间段划分的总体质量 M 定义为

$$M = \frac{L(t)}{w(t)} \quad (3)$$

给出了段内差异函数和段间差异函数的定义, 就可以计算出各自的复杂度。 $w(t)$ 的计算复杂度为 $O(\sum_i |L_i|) = O(n)$;

$L(t)$ 的计算复杂度为 $O(k^2)$ 。使段内差异最小和段间差异最大就作为划分告警序列时间段两个约束条件。

2 基于双约束的时间段划分算法

2.1 算法提出

算法首先随机选取 k 个时间点, 每个时间点都初始地代表了一个时间段的中心点。对剩余的每个告警事件根据其到各个时间段中心的距离, 将它赋给距离最近的时间段, 然后重新开始计算每个时间段的中心点。不断重复该过程, 直到准则函数收敛。定义准则函数为

$$E = \sum_{i=1}^k \sum_{t \in t_i} |t - t_{i/2}|^2 \quad (4)$$

其中: $t_{i/2}$ 表示时间段 t_i 的中间点; E 表示所有时间点的平方误差总和。这个准则就是为了满足段内差异函数和段间差异函数的要求, 使得每个划分的时间段尽可能紧凑和独立。

算法1 基于双约束的时间段划分算法

输入: 时间窗划分的个数 k 和包含 n 个告警事件的数据库。

输出: 满足平方误差总和最小的 k 个时间段。

(1) Select initial values for the middle points.

// 为 k 个时间段随机选取中间点

(2) for ($j = 1; j < = n; j++$) do {

(3) Assign each t_j to the time-window which has the closest mean; }

// 根据时间段中的中心点, 将其赋给最近的时间段内

(4) for ($i = 1; i < = k; i++$) do {

(5) find the middle point $t_{i/2}$ }

// 更新每个时间段内的中心点

(6) compute $E = \sum_{i=1}^k \sum_{t \in t_i} |t - t_{i/2}|^2$;

// 计算准则函数 E

(7) Repeat until E conversed.

// 重复计算直到准则函数收敛, 即变化趋势不明显

2.2 算法分析

基于双约束的时间段划分算法的优点是简单、快速, 在处理大数据集的时候, 该算法是相对可伸缩和高效率的。计算出算法的复杂度为 $O(n \cdot k \cdot m)$, 其中, n 是所有时间点的数目, k 是划分的时间段的数目, m 是迭代的次数。因为通常有 $k < n$ 且 $m < n$, 所以此算法结果可以达到局部最优。

该算法没有将时间段的平均值作为参照点, 主要原因是选择平均值对孤立点是敏感的, 而孤立点和异常点存在的数量很少, 结果就会严重影响时间段的均值, 从而影响时间段划分的质量。算法选择了时间段的中点作为参照点, 不仅可以去除孤立点和异常点的影响, 而且还可以将其作为噪声数据去除。

2.3 时间窗宽度和滑动步长选择

通过基于双约束的时间段划分算法, 可以将告警序列划分为多个相对集中而独立的告警子序列段。对于每个时间段, 设置各自的时间窗宽度和滑动步长。告警窗宽度的选择应该依据加权关联规则的挖掘效率、产生的有效加权关联规则的个数等因素来综合考虑, 窗口宽度 W 的取值范围为 $P_{\max} \leq W \leq \Delta W$ 。其中, $P_{\max}$ 为时间段内不同情景最大间隔的最大值, ΔW 为时间段的宽度。

滑动步长的选择以保证相邻窗口有足够的重叠为原则。滑动步长越大, 相邻窗口内的告警重叠得越多, 两个窗口内告

警的相关性也就越强,挖掘出的规则也就越多;相反,滑动步长越小,则会使相邻的两个窗口内告警重叠变少,挖掘的关联规则也越少。滑动步长的取值范围为 $Q_{\min} \leq s \leq \Delta W$ 。其中, $Q_{\min}$ 为时间段内不同情景最小间隔的最小值。

通过双约束的时间段划分,可以实现时间窗宽度和滑动步长根据告警事件的具体情况而动态变化,从而更加灵活地提取告警事务。

3 实验测试

实验采用的测试数据是某地区电信公司的告警记录,共采集了一周的告警数据,告警发生的时间是从2009年3月16日到2009年3月22日,共包括228 752条告警记录。主要测试基于双约束的时间窗划分算法和一般的基于均匀的时间窗划分算法在提取告警事务时的效率比较。

首先清洗原始告警数据,提取设备名称、设备所在的地址、告警类型、告警时间四个重要属性来表示告警事件。采用基于双约束的时间窗划分算法来提取告警事务,设定时间窗的数目分别为1000、2000、3000、4000和5000个,通过对原始告警的提取,得到的告警事务数目和提取的时间如表1所示。

表1 基于双约束的时间窗划分算法提取告警事务

时间窗数目	提取时间/s	告警事务数目	时间窗数目	提取时间/s	告警事务数目
1 000	65	25 762	4 000	99	28 365
2 000	72	26 450	5 000	123	29 043
3 000	87	27 745			

从表1可以看到,告警事务的数目是随着划分时间窗口的数目增加而增加的。然而,时间窗的数目越多,提取告警事务所需要的时间就越多,不能靠牺牲时间代价来换取告警事务数目的增加,所以要选择合适的时间窗数目。

然后比较了基于双约束的时间窗划分算法和基于均匀时间窗划分算法在提取告警事务时的效率,分别测试了提取1000~9000条告警事务所用的时间,图1给出了对比的结果。从图1中可以看出,本文提出的基于双约束的时间窗划分算法的提取效率明显高于基于均匀时间窗划分算法。两种算法的运行时间差别在提取告警事务数目达到5000、6000时达到最大,随着提取告警事务数目的增加,差别却逐渐减少。从实验的结果可以得出,要选择适合的时间窗数目,可以使提取告警事务的效率达到最高。

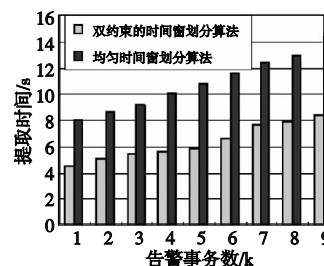


图1 采用两种不同方法提取告警事务的效率比较

4 结束语

由于告警数据分布得不均匀,在使用滑动时间窗提取告警事务的时候要考虑到提取的效率问题。本文提出了一种基于双约束的滑动时间窗方法来提取告警事务,算法选择了时间段的中点作为参照点,不仅可以提高提取的效率,还可以将孤立点和异常点作为噪声数据去除。实验证明该算法在提取告警事务时比传统的均匀时间窗方法效率更高。

参考文献:

- [1] VOSS A. Case reusing systems: survey, framework and guidelines [J]. *The Knowledge Engineering Review*, 1997, 12(1): 59-89.
- [2] MANNILA H, TOIVONEN H, VERKAMO A I. Discovery of frequent episodes in event sequences [J]. *Data Mining and Knowledge Discovery*, 2004, 1(3): 259-289.
- [3] AMANI N, FATHI M, DEHGHAN M. A case-based reasoning method for alarm filtering and correlation in telecommunication networks [C]//Proc of Canadian Conference on Electrical and Computer Engineering. 2005: 2182-2186.
- [4] KLEMETTINEN M, MANNILA H, TOIVONEN H. Rule discovery in telecommunication alarm data [J]. *Journal of Network and Systems Management*, 1999, 7(4): 395-423.
- [5] 王仲佳. 具有动态加权特性的关联规则算法及其在电信故障告警序列模式发掘中的应用[D]. 长春: 吉林大学, 2005.
- [6] HAUPTMANN M, LUBIN J H, ROSENBERG P, et al. The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer risk [J]. *Statistics in Medicine*, 2000, 19(1): 2185-2194.
- [7] Van VAERENBERGH, VIA J, SANTAMANA I. A sliding-window kernel RLS algorithm and its application to nonlinear channel identification [C]//Proc of International Conference on Acoustics, Speech and Signal Processing. 2006: 789-792.
- [8] Van VAERENBERGH S, VIA J, SANTAMANA I. Nonlinear system identification using a new sliding-window kernel RLS algorithm [J]. *Journal of Communications*, 2007, 2(3): 1-7.
- [5] ZHANG Zhong-ke, WANG Yun, LI Feng. Reliable packets delivery over segment-based multi-path in wireless Ad hoc networks [C] // Proc of the 6th International Conference on Pervasive Computing and Applications. 2011: 300-306.
- [6] SU S L, SU Yi-wen, CHEN C S, et al. A novel design of node-disjoint multi-path QoS routing in multi-hop Ad hoc networks [C]//Proc of the 41st International Conference on Dependable Systems and Networks Workshops. 2011: 292-296.
- [7] WANG Qing-hui, LU Yong-jie. A location-based multi-path routing approach in Ad hoc networks [C] //Proc of Chinese Control and Decision Conference. 2010: 2294-2296.
- [8] HANZO L, TAFAZOLLI R. QoS-aware routing and admission control in shadow-fading environments for multirate MANETs [J]. *IEEE Trans on Mobile Computing*, 2011, 36(4): 622-637.
- [9] VARAPRASAD G, WAHIDABANU R S D. Quality of service model for multimedia applications in a mobile Ad hoc network [J]. *IEEE Trans on Potentials*, 2011, 20(6): 44-47.
- [1] ZHANG Xin-ming, WANG En-bo, XIA Jing-jing, et al. An estimated distance-based routing protocol for mobile Ad hoc networks [J]. *IEEE Trans on Vehicular Technology*, 2011, 60(5): 3473-3484.
- [2] KATHIRAVELU T, SIVASUTHAN S. A hybrid reactive routing protocol for mobile Ad hoc networks [C]//Proc of the 6th IEEE International Conference on Industrial and Information Systems. 2011: 222-227.
- [3] DAI Bin, HE Jun-feng, WANG Hui, et al. iRoute: a scalable inter-domain multi-path routing framework for multimedia transmission [C]//Proc of International Conference on Multimedia Technology. 2011: 4966-4969.
- [4] QIN Dong-hong, YANG Jai-hai, ZHANG Bin, et al. Multipath inter-domain routing via deviation from primary path [C]//Proc of International Conference on Information Net-working. 2012: 222-227.

(上接第546页)

参考文献: