基于 Apriori 算法的时序关联关系数据挖掘装置的实现*

国悦婷 刘 磊 张 星
（华中科技大学自动化学院 武汉 430074）

摘 要 针对大规模网络频繁告警造成的短信网关压力和核心报警延迟甚至遗漏的问题，论文改进了 Apriori 算法用于告警合并，以适应运维场景的实际情况，并且实现了时序关联关系数据挖掘装置。该装置通过历史告警数据抓取、模型训练和数据验证测试三个步骤完成对告警数据的合并。与传统 Apriori 算法不同的是，针对时序告警序列规则一对一串行的特点，该优化算法省去了迭代过程并提出了一种新的置信度计算方式，解决了频繁告警项引起的置信度计算失真的问题，提高了关联规则的可信度。实验结果表明，该装置有效合并了告警信息，减轻了短信网关的压力，为海量告警信息故障的根因定位起到了积极的作用。

关键词 Apriori；时序关联关系；数据挖掘；置信度；告警合并
中图分类号 TP391 **DOI**：10. 3969/j. issn. 1672-9722. 2018. 02. 011

Realization of A Sequential Data Mining Device Based on Apriori Algorithm

GUO Yueting LIU Lei ZHANG Xing
（School of Automation，Huazhong University of Science and Technology，Wuhan 430074）

Abstract In view of the problems of the SMS Gateway pressure and the core alarm delay or omission caused by large-scale frequent network alarms，the improved Apriori algorithm is proposed for alarm merging to adapt to the actual operation scenarios and further the data mining device which is based on temporal correlation is realized. The device realizes the combination of alarm data through three steps：data capture，model training and data validation test. Contrary to the traditional Apriori algorithm，the improved algorithm removes the iteration process and proposes a new confidence level calculation method，according to the one-to-one characteristic of sequential alarm rules，solving the distortion problem of computing the confidence level caused by frequent alarms，as well as improving the credibility of association rules. Experimental results show that the device effectively merges the alarm information and reduce the pressure of SMS Gateway，which contributes to the root cause location of mass alarm failure information.

Key Words Apriori，sequential association，data mining，confidence level，alarm data merging
Class Number TP391

1 引言

随着互联网公司业务规模的快速增长，运维正面临着更为频繁的故障告警[1]。在大规模网络的运维工作中，每周海量的接警信息对运维工作人员而言工作量过大且难以及时查看，也会给短信网关带来较大压力，进一步影响故障的根因定位[2]，致使核心告警延迟甚至遗漏，造成重大损失[3]。因此，需要对海量告警信息进行告警收敛，以获取真正有价值的告警信息[4]。

目前，国内工业界针对以上问题，多采用的是告警合并策略[5]。首先，对告警进行分类，通常分为对业务有直接影响的告警[6]（如在线下降、服务器 ping 不通、业务进程崩溃、业务进程日志中出现

* 收稿日期：2017 年 8 月 9 日，修回日期：2017 年 9 月 15 日
作者简介：国悦婷，女，硕士研究生，研究方向：数据挖掘。刘磊，男，硕士研究生，研究方向：工业大数据与故障诊断。张星，男，硕士研究生，研究方向：大数据处理。

致命字符串等)和预警(如在线波动、CPU 和内存异常等)[7],其次,完成告警的去重、合并和聚合,最后,通过规则库过滤进行告警收敛[8],对于误报、连锁告警、大量告警等现象,会进行根因定位分析等,并做可视化数据呈现[9]。然而,采用简单的时间窗口合并、服务组织合并等方式的告警合并策略无法实现大量告警精准归类的目标,且合并效率较低[10]。

从历史告警数据中看,一个故障往往会引发很多相关策略的告警,这些告警在时序上存在一定的关联,所以我们可以从历史数据中挖掘不同策略间的时序关联关系,从而确定哪些规则是可以合并的,在告警的时候将关联策略合并展示,可以有效的减少告警条数、更清晰的接收故障信息[11]。本文从实际运维场景的需求出发,对 Apriori 算法作出相应改进,并在运维告警的时序信息场景中进行了实际应用,结果表明在告警合并的结果和效率上均获得了显著的改善效果。

2　问题定义

传统的关联关系挖掘方法使用支持度挖掘频繁项集,然后从频繁项集中产生规则,根据各规则置信度,给出关联规则。

定义 1　告警序列

告警序列 S 是由告警类型集合 E 上的多个有序的告警组成,表示为 $S=(s, T_s, T_e)$,s 为序列,T_s 为序列起始时间,T_e 为序列终止时间。例如在图 1 中,$T_s=0$,$T_e=600$,告警序列 s 由多个有序的告警向量(A,t)组成(其中 A 为告警监控项,t 为告警发生的时间,$A\in E$)。告警序列实例如图 1 所示。

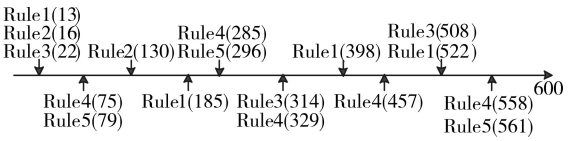

图 1　告警序列实例

定义 2　时间窗口

告警序列 $S=(s, T_s, T_e)$ 上的一个告警子项,可以表示成 $W=(w, t_s, t_e)$,$t_s<t<t_e$,其中 $w=t_e-t_s$ 称为时间窗口。

定义 3　告警串行关系

$\forall A_i(A_i\in E)$,$\forall A_j(A_j\in E)$,如果 $A_i\neq A_j$,$f:A_i\rightarrow A_j$ 或者 $f:A_j\rightarrow A_i$,则 A_i,A_j 之间是串行关系。告警情景 a 由告警事件 A 和 B 组成,且告警 A 会导致告警 B 出现,则 A 与 B 为串行关系。

3　Apriori 算法及改进

Apriori 算法是最有影响的挖掘布尔关联规则频繁项集的经典算法。

Apriori 算法中,使用的是一种称为逐层搜索的迭代方法,通过遍历数据库累计每个项的计数,并收集满足最小支持度的项,找出频繁项集的集合 L_1[12]。然后不断迭代直至不能找到频繁项集。其中,候选集产生的过程分为连接和剪枝两个部分。采用这种方式,使得所有的频繁项集既不会遗漏又不会重复。为提高频繁项集逐层产生的效率,Apriori 算法利用了一种称为先验性质的重要性质用于压缩搜索空间[13]。Apriori 算法流程如图 2 所示。

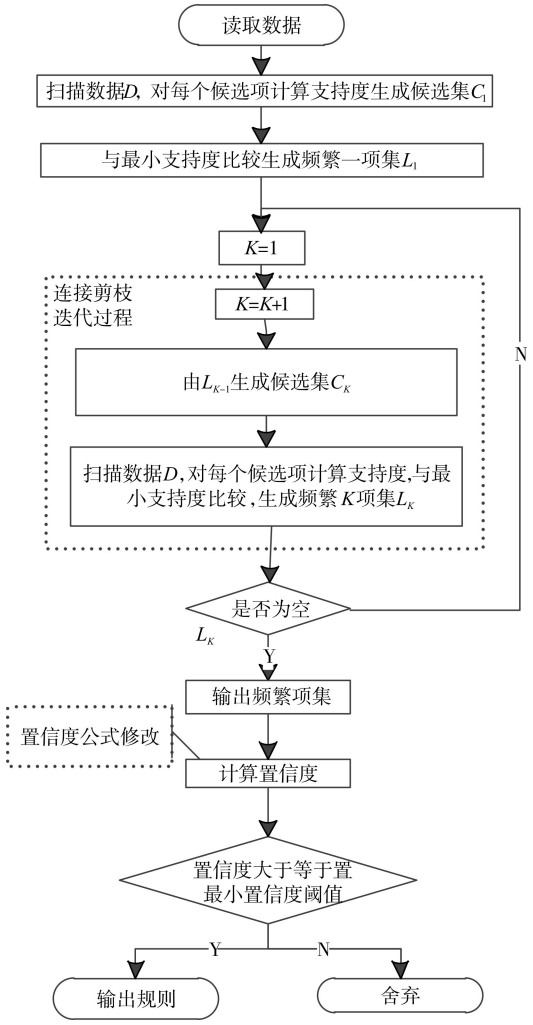

图 2　Apriori 算法流程图

先验性质　频繁项集的所有非空子集也一定是频繁的。

告警序列 $S=\{S_1, S_2, \cdots, S_m\}$ 是告警子序列的集合,设告警类型集合的数据 D 是数据库事务的集合,其中每个告警规则 R 是一个非空项集,使得 R

$\subseteq S$。每一个规则都有一个标识符,设为RID。设A是一个告警规则项集,规则R包含A,当且仅当$A \subseteq R$。关联规则是形如$A=>B$的蕴含式,其中$A \subset S$,$B \subset S$,$A \neq \varnothing$,$B \neq \varnothing$,并且$A \cap B=\varnothing$。规则$A=>B$在事务集D中成立,具有支持度s,其中s是D中事务包含$A \cup B$的百分比[14]。它是概率$P(A \cup B)$。规则$A \rightarrow B$在事务集D中具有置信度c,其中c是D中包含A的事务同时也包含B的事务的百分比。式(1)和(2)给出了支持度和置信度的定义式[15]。

$$\text{支持度} \quad \text{support}(A \rightarrow B) = P(A \cup B) \tag{1}$$

$$\text{置信度} \quad \text{confidence}(A \rightarrow B) = P(B|A) = \frac{\text{support}\left(A \bigcup B\right)}{\text{support}(A)} = \frac{\text{support\_count}\left(A \bigcup B\right)}{\text{support\_count}(A)} \tag{2}$$

在运维场景中,告警在时序上存在一定的关联,如Rule1 → Rule2,呈现一对一串行的特点。因此可对Apriori算法进行简化,去掉迭代过程以提升运行效率,其过程是先建立带时间窗口的支持度表,然后计算得出计数表,最后在得到置信度表的时候,对置信度的计算做调整,避免频繁报警项作为分母时引起置信度值失真。因此将置信度的公式调整为式(3)。

$$\text{新置信度} \quad confidence(A \rightarrow B) = \frac{\sigma(A \cap B)}{\sigma(A \bigcup B)} \tag{3}$$

由 Card($A \bigcup B$) =Card(A) + Card(B) −Card($A \cap B$),置信度公式在矩阵计算中可记为式(4):

$$confidence(A \rightarrow B) = \frac{\sigma(A \cap B)}{\sigma(A) + \sigma(B) - \sigma(A \cap B)} \tag{4}$$

如果项集S的置信度满足预定义的最小置信度阈值min_conf,如满足式(5):

$$confidence(A \rightarrow B) \geq \min\_conf \tag{5}$$

又因为规则由频繁项集产生,因此每个规则都自动地满足最小支持度。同时满足最小支持度阈值和最小置信度阈值的规则称为强规则,则S是频繁项集,规则$A \rightarrow B$为强规则。

4 实验内容

为达到对告警信息合并收敛的效果,设计并实现了数据挖掘装置,其整体框架分为三部分:抓取模块、模型训练模块和测试评估模块。该框架支持历史数据的爬取存储、挖掘训练和测试,并将三个模块的代码解耦,并具有高内聚、低耦合的特点。

数据抓取部分从各个存储系统中抓取告警数据,经过数据清洗存入本地磁盘,模型训练部分使用相应的策略训练数据模型,得到挖掘规则,模型测试部分根据本地磁盘的数据和挖掘规则,得到测试结果。系统结构如图3所示。

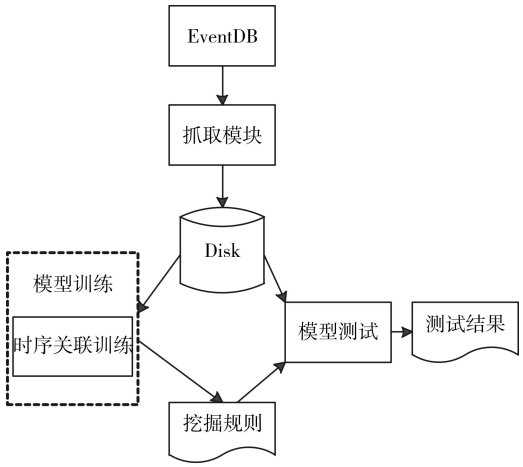

图3 数据挖掘装置框架图

实验过程共分为三个步骤:1)告警历史数据抓取;2)挖掘算法训练;3)测试评估结果。

第一步,需对历史数据进行抓取,该过程包括从各个存储系统中抓取离线数据和将清洗后的数据存入本地磁盘两个步骤。以图1的告警序列为例,采用60s的时间窗口,将抓取的告警数据整理为表1。

表1 带时间窗口的告警序列

W序号	时间窗口W	告警监控项
1	(60,0,60)	Rule1,Rule2,Rule3
2	(60,60,120)	Rule4,Rule5
3	(60,120,180)	Rule2
4	(60,180,240)	Rule1
5	(60,240,300)	Rule4,Rule5
6	(60,300,360)	Rule3,Rule4
7	(60,360,420)	Rule1
8	(60,420,480)	Rule4
9	(60,480,540)	Rule3,Rule1
10	(60,540,600)	Rule4,Rule5

对时序数据流的数据挖掘采用约定起始时间和固定时间间隔的滑动窗口,规定时间窗口为1个时间单位,以Rule1 → Rule4为例,W1⇒W2,W4⇒W5,W7⇒W8,W9⇒W10这四个相邻的时间窗口均支持Rule1 → Rule4的关联关系,因此Rule1 → Rule4的支持度计数为4。依此规则,计算所得如表2所示。

第二步,对时序关系进行模型训练。关联关系的挖掘包含两个过程:1)找出所有的频繁项集;2)由频繁项集产生强关联(强规则)。

根据步骤一,遍历表1统计每个规则出现的次数,得出所有规则的支持度计数,如表3所示。

表 2　带时间窗口的支持度计数表

	Rule1	Rule2	Rule3	Rule4	Rule5
Rule1	–	1	1	4	3
Rule2	2	–	1	1	1
Rule3	3	1	–	3	2
Rule4	1	1	2	–	3
Rule5	0	1	1	3	–

表 3　支持度计数

规则	次数 $\sigma(Rule)$
Rule1	4
Rule 2	2
Rule 3	3
Rule 4	4
Rule 5	3

由式(4),得出置信度表如表 4 所示。

表 4　置信度

	Rule1	Rule2	Rule3	Rule4	Rule5
Rule1	–	0.2	0.16	1	0.75
Rule2	0.5	–	0.25	0.2	0.25
Rule3	0.16	0.25	–	0.75	0.33
Rule4	0.14	0.2	0.6	–	0.75
Rule5	0	0.25	0.2	0.75	–

对于 S 的每个非空子集 s,如果

$$\frac{\sigma(A\cap B)}{\sigma(A)+\sigma(B)-\sigma(A\cap B)}\geq \text{min\_conf},$$

则输出规则 $A\Rightarrow B$。

由经验 min_conf 设置为 0.3,则强规则为 Rule1 → Rule5, Rule2 → Rule1, Rule3 → Rule4, Rule3 → Rule5, Rule4 → Rule3, Rule4 → Rule5, Rule5 → Rule4。

第三步,根据训练得到的关联关系进行数据测试。当遇到有关联关系的告警项时,将其收敛为同一条告警信息,并计算出收敛率。

5　结果及分析

本文采用 python 编程实现算法,操作系统为 Linux,配置 CPU 频率为 2400.084MHz。实验中采用某公司网管数据库连续一个月的原始告警数据做为测试数据,大约 200 万条告警记录。

将训练数据的时间设定为 2016 年 9 月 1 日到 2016 年 10 月 1 日,并将测试数据时间设定为 2016 年 11 月 1 日到 2016 年 12 月 1 日,输出如表 5。

表 5　告警收敛结果

最小置信度阈值	最小支持度阈值	合并后报警数量	初始报警数量	合并率
0.3	5	1748688	2047217	85.42%

结果显示,通过算法的合并使得原始告警数量有了较大程度的削减。功能上完成了对有关联关系的规则的合并,起到了告警收敛的作用,减少了冗余告警数量。对于最小自信度阈值的选取,既不能过小,不然会使数据关联度增大,本来没有关联的数据进行合并后,对告警信息报警的准确率有一定干扰,也不能选择太大,不然会削弱告警合并的效果,无法达到理想的合并效果。

6　结语

本文通过对 Apriori 算法的改进和置信度公式的优化,设计并实现了时序关联关系数据挖掘装置,并将其用于大规模网络告警合并场景,从而减少了冗余告警数量,减轻了短信网关的压力,为海量告警信息故障的根因定位起到了积极的作用,减少了核心告警延迟和遗漏的情况。根据告警场景的实际情况,改进后的 Apriori 算法能更好地挖掘出关联关系规则,进而提高了该算法的实用性。本文的方法对已发生的告警起到了有效的合并作用,下一步的工作是对告警数量进行预测,对未来预期时间的告警数量进行预测,当超过该预测值的时候,则认为疑似出现大规模告警,然后利用本文的方法对告警信息进行合并展示。

参 考 文 献

[1] Dattatraya V K, Shaila D A. A Universal Object Oriented Expert System Frame Work for Fault Diagnosis[J]. International Journal of Intelligence Science, 2012, 2(3): 63-70.

[2] 李彤岩. 基于数据挖掘的通信网告警相关性分析研究[D]. 成都:电子科技大学,2010.
LI Tongyan. Research on alarm correlation analysis of communication network based on Data Mining[D]. Chengdu: University of Electronic Science and technology, 2010.

[3] 邓歆,孟洛明. 基于贝叶斯学习的告警相关性分析[J]. 计算机工程,2007,33(12):40-42.
DENG Xin, MENG Luoming. Alarm correlation analysis based on Bayesian learning[J]. Computer Engineering, 2007,33(12):40-42.

[4] 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[J]. 计算机研究与发展,2009,46(8):1304-1315.
TIAN Zhihong, ZHANG Yongzheng, ZHANG Weizhe, et al. Adaptive alarm correlation based on pattern mining and clustering analysis[J]. Journal of Computer research and

(下转第 269 页)

tern matching based on large scale corpus[J]. Journal of Chinese Information Processing,2007,21(5):31-35.
[5] 蒋德良. 基于规则匹配的突发事件结果信息抽取研究[J]. 计算机工程与设计,2010,31(14):3294-3297.
JIANG Deliang. Research on rule matching based information extraction for unexpected events[J]. Computer engineering and design,2010,31(14):3294-3297.
[6] 王昀,苑春法. 基于转换的时间—事件关系映射[J]. 中文信息学报,2004,18(4):23-30.
WANG Yun,YUAN Chunfa. Time event relationship mapping based on transformation[J]. Journal of Chinese Information Processing,2004,18(4):23-30.
[7] 李文捷,周明. 基于语料库的中文最长名词短语的自动提取[J]. 计算语言学进展与应用,1995:119-124.
LI Wenjie, ZHOU Ming. Corpus based automatic extraction of Chinese longest noun phrases[J]. Advances and applications in Computational Linguistics,1995:119-124.
[8] Kiyoshi Sudo 2004. Unsupervised Diseovery of Extraction Patterns for Information Extraetion [D]. Department of Computer Science.New York University,September,2004.
[9] GB18218-2000, 重大危险源辨识[S]
GB18218-2000, Identification of major hazard installations[S].
[10] 孙宏林,俞士坟. 浅层句法分析方法概述[J]. 当代语育学,2000,2(2):74-83.
SUN Honglin, YU Shiwen. Overview of shallow parsing methods [J]. Contemporary language education,2000,2(2):74-83.
[11] 辛宵,范士喜,王轩,等. 基于最大熵的依存句法分析[J]. 中文信息学报,2009,23(2):18-22.
XIN Xiao, FAN Shixi, WANG Xuan, et al. Dependency parsing based on maximum entropy[J]. Journal of Chinese Information Processing,2009,23(2):18-22.
[12] 李洪波,陈军. Prim 最小生成树算法的动态优化[J]. 计算机工程与应用,2007,43(12):69-73.
LI Hongbo, CHEN Jun. Dynamic optimization of Prim minimum spanning tree algorithm [J]. Computer engineering and Applications,2007,43(12):69-73.
[13] 夏天. 词语位置加权 TextRank 的关键词抽取研究[J]. 现代图书情报技术,2013(9):30-34.
XIA Tian. Keyword extraction of word position weighted TextRank [J]. New Technology of Library and Information Service,2013(9):30-34.

(上接第 263 页)

development,2009,46(8):1304-1315.
[5] J.Han,M.Kamer. 数据挖掘概念与技术[M]. 2 版. 北京:机械工业出版社,2007.
J.Han, M.Kamer. Data mining concepts and techniques [M]. Second Edition,Beijing:China machine press,2007.
[6] Unil Y. Efficient Mining of Weighted Interesting Patterns with a Strong Weight and/or Support Affinity[J]. Information Sciences,2007,177(17):3477-3499.
[7] 蔡伟杰,张晓辉,朱建秋,等. 关联规则挖掘综述[J]. 计算机工程,2001,(05):31-33,49.
CAI Weijie,ZHANG Xiaohui,ZHU Jianqiu, et al. Survey of Association Rule Generation [J]. Computer Engineering,2001,(05):31-33,49.
[8] 吴扬扬,陈怀南. 基于关联规则的通信网络告警相关性分析模型[J]. 通讯和计算机,2004,12(1):57-63.
WU Yangyang, CHEN Huainan. Alarm correlation analysis model of communication network based on association rules[J]. Journal of Communication and Computer,2004,12(1):57-63.
[9] 夏海涛,詹志强. 新一代网络管理技术[J]. 北京邮电大学出版社,2003,05.
XIA Haitao, ZHAN Zhiqiang. New generation network management technology [J]. Beijing University of Posts and Telecommunications Press, 2003,05.
[10] 杨芬. 基于约束的关联规则挖掘[D]. 武汉:华中科技大学,2004.
YANG Fen. Association Rules Mining Based on Constraints [D]. Wuhan: Huazhong University of Science and Technology,2004.
[11] Pang-Ning Tan Michael Steinbach Vipin Kumar. Introduction to Data Mining[M]. March 25, 2006.
[12] Zhen Yun Liao, Xiu Fen Fu, Ya Guang Wang. The Research of Improved Apriori Algorithm [J]. Applied Mechanics and Materials,2013,2171(263).
[13] Gao H S, Li Y M.An Efficient Communication Network SDH Alarm Association Rule Mining Algorithm[J]. Advanced Materials Research, 2014.
[14] N. Badal, Shruti Tripathi. Frequent Data Itemset Mining Using VS_Apriori Algorithms [J]. International Journal on Computer Science and Engineering,2010,2(4).
[15] Li Rong Tong, Jun Zhang, Lei Ma, Li Xin, Shuang Hu, Jian Feng He. An Improved Apriori Algorithm Based on LinkedList for the Prevention of Clinic Pharmaceutical Conflict [J]. Applied Mechanics and Materials, 2014, 2987(513).