基于关键属性约束的关联规则挖掘在日志分析中的应用

金可仲

（温州大学计算机科学与工程学院，浙江温州 325035）

摘 要：日志是计算机取证、入侵检测分析的重要数据来源，运用关联规则挖掘算法对日志进行分析是获取日志中所蕴含有用信息的重要方法. 针对基于置信度-支持度框架的常用关联规则挖掘算法在日志分析中存在的不足，引入日志关键属性的概念，提出了基于关键属性约束的关联规则挖掘算法. 实验结果表明，该算法能有效阻止无趣规则的产生，提高挖掘结果的有效性.

关键词：日志；关联规则挖掘；关键属性

中图分类号：TP311 **文献标识码**：A **文章编号**：1006-0375(2008)01-0056-05

为了维护自身资源的运行状况，计算机系统都会有相应的日志文件，来记录与系统或应用有关的日常管理事务信息或操作事务信息，文献[1]给出了日志的定义：所谓日志，就是指系统指定对象的某些操作和操作结果按时间排序的集合. 文献[2]给出了不同来源、不同格式类型日志信息的集中式统一管理方法. 收集、保存并管理日志，对日志进行分析，从中可得出一些有效的决策辅助信息，这对于重构系统安全性、辅助网络站点的设计和个性化处理等有重要意义.

由于日志具有数据量大、不易读懂的特点，其中所蕴含的有用信息难以发现. 将数据挖掘技术应用于日志分析是该领域当前的一个研究热点，如利用关联规则方法挖掘隐藏在日志记录之间的相互关系，利用序列模式分析方法寻找日志记录中的时间序列关系，利用异常点分析技术发现数据中的孤立现象等，其中关联规则算法的应用尤为广泛，包括 FP 算法、Tvp 算法、Apriori 算法等[3]. 本文针对基于置信度-支持度框架的常用关联规则挖掘算法在日志分析中存在的不足，引入日志关键属性的概念，提出了基于关键属性约束的关联规则挖掘算法. 实验结果表明，该算法能有效阻止无趣规则的产生，提高挖掘结果的有效性.

1 基于关键属性的关联规则挖掘算法

1.1 常用算法在日志分析中的局限性

倘若将常用的关联规则挖掘算法直接应用于日志数据的挖掘分析，会出现以下问题：

（1）产生许多不相关（或称为无趣）的规则，如对表 1 的网络日志数据进行关联规则挖掘，会产生这样一些关联规则：$src\_bytes = 200 \Rightarrow flag = SF$，由于主机发送的字节数与连接的状态没有直接关系，因此这样的规则是无用的规则，而且此类规则如果不排除，不仅不能为日志

收稿日期：2007-06-27

作者简介：金可仲(1979-)，男，浙江永嘉人，助教，硕士，研究方向：信息安全

分析提供有用的帮助，还会起到误导作用，淹没真正有价值的信息.

表 1 网络日志数据

service	localhost	remotehost	src_bytes	dest_bytes	state	flag
ftp-data	18	129.7.1.9	212	0	SF	L
nntp	4	128.49.4.103	90447	1200	SF	L
smtp	3	128.8.142.5	893	331	SF	-
ftp	18	129.7.1.9	1998	2450	SF	L
smtp	29	138.95.24.78	0	0	REJ	-
telnet	12	192.35.222.222	3	175	SF	L
…	…	…	…	…	…	…

（2）候选频繁集合的产生，需要多次扫描日志数据库，这极大地影响了算法的效率. 网络日志中不同字段的重要性是不同的，如果不加限制，全部扫描，显然很不合理.

（3）常用关联规则挖掘算法使用置信度-支持度框架，尽管这种框架体系可以排除一些小概率的无趣规则，但同样也会把一些小规模的重要规则排除在外.

因此，为了更好地在日志分析中应用关联规则挖掘算法，必须对算法加以改进，增加限制规则产生的条件，提高算法的执行效率和规则的有效性.

1.2 关键属性概念的提出

在日志记录中，各个属性的重要性是不同的，有的属性对描述数据起着关键作用，有的属性只提供辅助信息. 如在描述网络连接的日志记录中，每一条记录表示一个连接，一个网络连接可以被一个五元组唯一地标识，即<timestamp srcIP destIP srcPort destPort(service)>，这些属性可称为关键属性[4-5].

定义 1 设 U 是所要研究的属性的集合，F 是属性间依赖关系的集合，$K \subseteq U$，若 U 完全函数依赖于 K，则称 K 是关键属性集，其任一属性 $k \in K$ 称为关键属性.

那么，可以用关键属性作为衡量关联规则 r 是否有趣的标准，即

$$I(r) = f\ (I_A(r), s(r), c(r))$$

其中 I_A 有这样的性质：

if 模式 r 含有关键属性 *then*

I_A=1

else

I_A=0

含有这些关键属性的规则叫做“相关”规则，如果不含这些关键属性，只由非关键属性表示的规则是“不相关”规则.

2 基于关键属性的关联规则挖掘算法

根据上述关键属性的定义，在关联规则产生的过程中，不仅以支持度和置信度框架来衡量，同时加入关键属性的约束[6]，可以大大减少无趣规则的产生，即当算法产生一条新的规则时，要求其中至少包括一个关键属性.

基于关键属性约束的关联规则算法表示如下：

输入：日志记录数据库 D，关键属性集合 K，

相关参数（最小支持度 s，最小置信度 c）

输出：包含关键属性的关联规则集合 R

(1) $L_1 = \{\text{频繁}1-\text{数据项集}\}$；//通过搜索数据库生成

(2) $\text{for}(k = 2; L_{K-1} \neq \varphi; k++)$

(3) $C_K = \text{apriori\_gen}(L_{K-1})$；//产生新的侯选集

(4) 对每一日志记录 $T \in D$

(5) $C_T = \text{subset}(C_K, T)$；//事件 T 中包含的侯选集

(6) 对每一个元素 $c \in C_T$

(7) $c.\text{count}++$；

(8) $L_K = \{c \in C_K | c.\text{count} \geq s\}$；

(9) end

(10) 对于每个频繁 k-项集 l，如果有 $l \bigcap R \neq \phi$，则生成其所有非空子集 α

(11) 对每个非空子集 α，如果有 $\dfrac{\text{support\_count}(l)}{\text{support\_count}(\alpha)} \geq c$

(12) 把规则 $\alpha \Rightarrow (l\text{-}\alpha)$, c, s 加入 R

(13) 返回规则集合 R

3 实验比较和结果分析

以来自 http://ita.ee.lbl.gov/html/traces.html 的日志记录数据文件 lbl-conn-7.tar.Z 中的 TCP 连接日志记录(Lawrence Berkeley Laboratory 与外界之间 30 天的日志记录，截取其中部分数据，约 5000 条记录)为数据源，对采用常用关联挖掘算法和采用关键属性约束关联规则挖掘算法的效果进行分析对比. 其中置信度设置为 95%，支持度设置从 1%到 10%. 作为约束项的关键属性包括 protocol、localhost、remotehost. 选取以上三项为属性约束项的目的，是为了使产生的规则包含网络日志中的通信双方用户和服务，以反映用户在网络连接中的行为. 结果如图 1 所示：

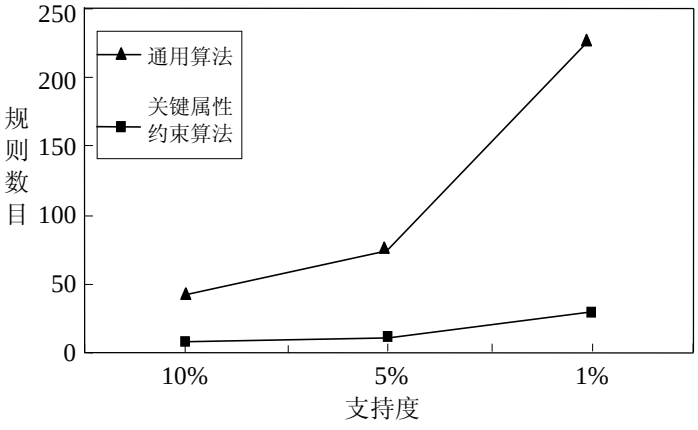

图 1　关键属性约束对产生规则数目的影响

从图 1 可以看出，随着支持度阈值的下降，是否引入关键属性进行约束挖掘，对产生规则的数目影响很大．因为日志数据往往是海量的，借助数据挖掘的手段，是为了提取其中有用的关键信息，支持日志分析工作．如果产生的结果信息量很大，但是无用规则太多，反而会扰乱日志分析工作，所以关键属性约束的引入，能有效阻止无趣规则的产生，有助于挖掘到一些有用的规则信息．

对挖掘得到的规则，经过分析得出如下结论：

（1）通用算法产生的数目虽然多，但是包含大量的关于时间、发送字节数、日期、状态、标志等关联的无用规则，如

$$state = SF, localhost = 12 \Rightarrow flag = L$$

这些规则只是算法产生的一些副产品，无法为日志分析提供有用的线索，反而使可能有用的信息淹没在其中，增加了分析的难度，降低了效率．

（2）基于属性约束的挖掘算法得到的结果，虽然数量不多，但是都是包含日志中重要属性的规则，反映了不同属性之间存在的事实上的一些联系，如

$$protocol = telnet, remotehost = 192.35.222.222 \Rightarrow localhost = 12$$

通过对原始日志的仔细分析发现，能找到 675 条关于远程主机 192.35.222.222 和本地主机 12 之间进行 telnet 通信的记录，这表明规则反映了日志记录中的事实情况．

实验结果显示，基于关键属性约束的关联规则挖掘算法，与通用算法相比，在日志分析中更加有效．

4 结　语

日志分析的主要目标，是从海量的日志数据中找出隐含的关联信息，为计算机入侵检测、取证分析等提供有力支持．如何对日志进行有效的分析是当前该领域重要的研究课题．本文分析了关联规则挖掘算法在日志分析中的应用，引入日志关键属性的概念，提出了基于关键属性约束的关联规则挖掘算法．实验结果表明，该算法的结果在有效性方面要优于通用关联规则挖掘算法的结果．

参考文献

[1] 林晓东, 刘心松. 文件系统中日志技术的研究[J]. 计算机应用, 1998, 18(1): 30-32.

[2] 赵小敏, 侯强, 陈庆章. 系统日志的安全管理方案和分析处理策略[J]. 计算机工程与科学, 2003, 25(3): 44-47.

[3] 文娟, 薛永生, 段江娇, 等. 基于关联规则的日志分析系统的设计与实现[J]. 厦门大学学报: 自然科学版, 2003, 44: 258-261.

[4] 徐菁. 基于数据挖掘技术的入侵监测模型[D]. 北京: 中国科学院高能物理研究所, 2001: 70-71.

[5] Lee W, Stolfo S J, Mok K. Algorithms for Mining System Audit Data[C] // Lin T Y, Yao Y Y, Zadeh L A, et al. Data Mining, Rough Sets, and Granular Computing. Heidelberg: Physica-Verlag, 2002: 166-189.

[6] Srikant R, Vu Q. Mining association rules with Item Constraints[C] // Heckerman D, Mannila H, Pregibon D, et al. Proceedings of the 3rd International Conference on Data Mining and Knowledge Discovery. California: AAAI Press, 1997: 67-73.

Application on Log Analysis Using Association Rule Mining Algorithm with Key Item Constraint

JIN Kezhong
(School of Computer Science and Engineering, Wenzhou University, Wenzhou, China 325035)

Abstract: Computer log is important data source for computer forensic analysis and intrusion detection analysis. Mining algorithms of association rule are important methods to find implicative useful information. But common mining algorithms of association rule which based on the frame of confidence and support are unfit in log analysis. To solve this problem, this paper introduces key attributes of log and puts forward an algorithm with key attributes constraints to mine association rules. The result of experiment shows that the algorithm can reduce the number of domain-independent rules and improve the validity of mining result.

Key words: Log; Mining of association rule; Key attribute

（编辑：王一芳）