doi:10.3969/j.issn.1002-0802.2014.08.017

基于关联规则挖掘的智能云防护技术研究*

方忠进[1,2,3],夏志华[1,3],周 舒[1,3]

(1.南京信息工程大学 计算机与软件学院,江苏 南京 210044;
2.南京信息工程大学 滨江学院,江苏 南京 210044;
3.南京信息工程大学 江苏省网络监控工程中心,江苏 南京 210044)

摘 要: 针对传统安全防护技术存在的被动防御、效率较低的缺陷,提出了一种基于关联规则挖掘的智能云防护技术。该技术引入一种改进的FP-Growth挖掘算法,用来提取数据的特征信息,提交到云端,通过将此特征信息与入侵特征库和网络正常活动行为特征库进行匹配,从而发现病毒和攻击行为。特征库根据庞大的云探针系统采集的数据样本不断更新完善。实验结果表明,基于改进的FP-Growth挖掘算法的智能云防护技术对攻击行为响应较快,检测效率较高,准确记录攻击日志,具有较好的安全防护性能。

关键词: 关联规则 云安全 FP-Growth算法 数据挖掘
中图分类号: TP391.4 **文献标志码:** A **文章编号:** 1002-0802(2014)08-0925-05

Study on the Intelligent Cloud Protection Technology based on Association Rule Mining

FANG Zhong-jin[1,2,3], XIA Zhi-hua[1,3], ZHOU Shu[1,3]

(1. School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing Jiangsu 210044, China;
2. Binjiang College, Nanjing University of Information Science & Technology, Nanjing Jiangsu 210044, China;
3. Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing Jiangsu 210044, China)

Abstract: According to the limit of passive defense and low detection efficiency of traditional security technology, an intelligent cloud protection technology based on association rule mining is proposed. An improved FP-Growth algorithm is applied to extract the feature information of data. The feature information is matched with intrusion feature database and normal activities feature database to detect viruses and attacks. Feature database is continuously updated and improved according to the data samples collected by the huge cloud probe system. Experimental results show that the intelligent cloud protection technology based on the improved FP-Growth algorithm can detect attacks in time, and record attack log accurately. It has a high detection efficiency and good protection performance.

Key words: association rule; cloud security; FP-Growth algorithm; data mining

0 引 言

随着网络技术的快速发展,网络攻击、数据泄露等信息安全事件频发,防火墙等传统的静态安全防御技术对于攻击缺乏主动的响应,已无法满足日益复杂的安全应用要求。利用云计算来提高安全防护能力已成为信息安全领域研究的新课题。“云计算”直接起源于2007年Amazon EC2产品和Google-IBM分布式计算项目,这两个项目直接使用了“云计算”这一概念[1]。云计算是网格计算、分布式计算、并行计算、效用计算、网格存储、虚拟化、负载均衡等传统计算机和网络技术发展融合的产

* 收稿日期:2014-05-06;修回日期:2014-06-24 Received date:2014-05-06;Revised date:2014-06-24

物，它能有效解决网格计算无法同时支持异构多任务体系、无法实现资源动态流转的不足[2-3]。云计算以新的业务模式提供高性能、低成本的持续计算和存储服务，支撑各类信息化应用。

针对云计算具有的资源动态流转、支持海量信息处理的特点，提出一种智能云防护技术。云平台的每一个客户端作为一个探针，网状的大量探针提取出数据信息中的特征，提交给云端进行分析和处理，对于木马、病毒等攻击行为，云端将解决方案下发给探针。当某个探针遭受新的安全威胁，智能云防护技术将会获取其特征信息，更新入侵特征库。最新的入侵、病毒特征信息会及时更新共享，并适时给出解决方案。通过分布在全球各地的探针，云端数据库积累了大量的特征信息，及其全面、准确地创建了网络正常活动行为特征库，利用特征库来监视是否存在安全威胁行为，正常行为特征库精确地提前定义出可能发生的安全风险与威胁。

数据挖掘，也称数据库中的知识发现（KDD，Knowledge Discovery in Database），指从大型数据库或数据仓库中提取人们感兴趣的知识，这些知识是隐含的、事先未知的潜在有用信息，提取的知识一般可表示为概念（Concepts）、规则（Rules）、规律（Regularities）、模式（Patterns）等形式[4]。简单地说，数据挖掘就是从大量的数据中抽取挖掘出未知的、有价值的模式或规律知识的复杂过程。

关联规则挖掘是数据挖掘中最活跃、最有效的研究方法之一，本文提出了一种基于关联规则挖掘的智能云防护技术框架，采用一种改进的 FP – Growth 算法进行网络数据特征提取，分析了相关实验结果。

1 技术架构

智能云防护技术将云端和客户端有机结合，构建高效的智能威胁收集系统，通过数据挖掘技术创建入侵特征库和正常行为特征库，云端根据各探针收集上报的数据信息，运用数据挖掘技术，对入侵特征库和正常行为特征库进行实时更新完善，及时发现新型威胁并将解决方案分发到各客户端，从而将安全威胁消灭在最初阶段。

基于关联规则挖掘的智能云防护技术架构如图 1所示。

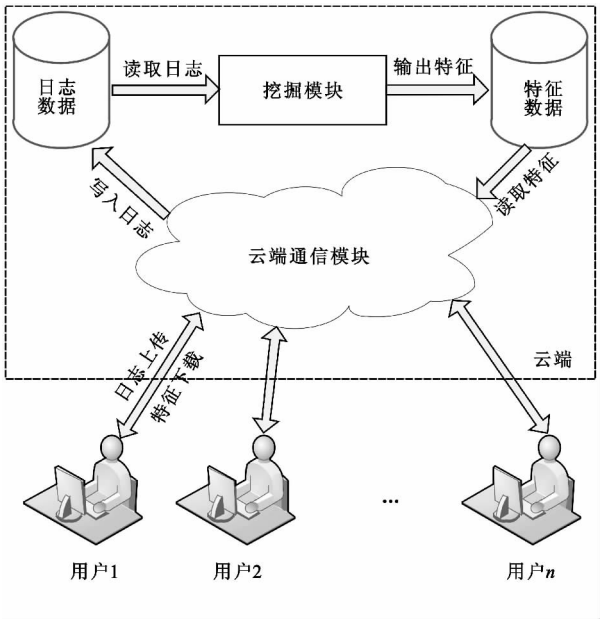

图 1 基于关联规则挖掘的智能云防护技术架构

Fig. 1 Structure ofthe intelligent cloud protection technology based on association rule mining

系统方案设计如下：

1）位于不同位置的用户上传日志文件到云端通信模块。

2）云端通信模块将日志文件写入到特定的数据库。

3）挖掘模块定时读取日志数据库并执行数据挖掘操作，输出特征内容写入到特征数据库

4）用户端定期通过云端通信模块访问特征数据库，更新本地特征库。用户端根据特征库拦截可能存在的攻击行为。

建立特征库的过程见图 2。

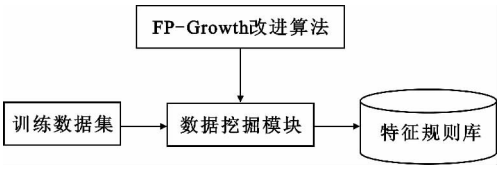

图 2 应用关联规则挖掘建立特征库的过程

Fig. 2 Building feature database with application of association rule mining

2 改进的 FP – Growth 关联算法

数据挖掘是一种有效地针对海量数据进行分析处理的技术，能够提取出大量网络数据中潜在的规则信息，本文使用中的数据挖掘模块采用的是关联分析算法，关联规则挖掘可以发现数据中项集之间有用的关联或相关联系，对于安全防护系统而言，则

可以提取出网络数据中的特征模式[5]。本文采用一种改进的 FP – Growth 关联算法。

2.1　FP – Growth 关联算法分析

FP – Growth (Frequent – pattern Growth) 算法是由 Jiawei Han，Jian Pei 等人提出的，该算法采用模式增长的方法，从大规模数据中挖掘出频繁模式，这样就不需要额外产生候选集。该算法使用的策略如下: 算法将数据库中的频繁项集合进行压缩，使之成为一棵频繁模式树(FP – tree) 的同时，关联信息依旧被保留。将压缩后的数据库分成一组条件数据库，每个条件数据关联一个频繁项，进而分别挖掘这些条件数据库。该算法可以通过递归方式发现一些短模式，替代了原本复杂地发现长频繁模式的问题，然后连接后缀，效率上较之经典关联 Apriori 算法有了很大的提高[4,6]。

FP – tree 的定义如下:

1) 树有一个根结点，用空值“null”来标记，它的子树是一个项前缀子树的集合。除此之外，还有一张表来存贮每个频繁项结点的头结点。

2) 项前缀子树各节点由以下 3 部分组成: 项目名称、频繁项计数器和一个结点指针。其中项目名称表示的是该结点所描述的频繁项，频繁项计数器记录的是从根结点出发到此结点的详细路径所包含的事务的数量，结点指针所指向的是项前缀子树里下面一个和此结点具有同样名称的结点，如果后续已经没有表示该项目的其他结点了，则设这个结点的指针值为空值。

3) 频繁项头表中的各个项由两个字段构成: 项目名称与头指针。头指针指向项前缀子树中第一个同名项。

FP – Growth 算法的主要步骤[7] 如下:

1) 扫描事务数据库一遍得到各项目的频度，根据最小支持度得到频繁项; 对频繁项按其频度由大到小排列成表 L，形成头表。

2) 再次扫描事物数据库一遍，对每一条交易中的所有频繁项，按表 L 中的次序插入到 FP – Tree 中。

3) 调用 FP_growth 算法对 FP – Tree 进行挖掘。

FP – Growth 算法在本质上不同于 Apriori 算法的候选生成 – 筛选方法，克服了 Apriori 算法的缺点，通过采用关联规则挖掘新思路，解决了传统方法中需要产生候选项集的问题。该算法中所构造出的 FP – tree 是一种具有高压缩度的数据结构，存储的是与频繁模式相关的重要内容; 此算法仅对数据库扫描两遍，将扫描时间减少到最少，提高了效率。数据挖掘的主要工作就是对累加值进行计算并对前缀树进行调整，这种在存储和计算资源上的花费要比 Apriori 算法中使用的候选项集产生算法与模式匹配操作算法小得多。

2.2　FP – Growth 改进算法

尽管 FP – Growth 算法有不少优点，但它仍有不足之处。例如在对频繁模式进行挖掘时，其需要使用递归算法不断地生成条件 FP – tree，当生成一个频繁模式时就会产生一个与其对应的条件 FP – tree。在最小支持度相对较小时，即使挖掘不太大的数据库，也将产生成千上万的条件 FP – tree。如此多的条件 FP – tree 的动态创建和释放，会耗费非常多的 CPU 处理时间和存储空间，这对挖掘效率的影响是非常大的。同时，FP – tree 与条件 FP – tree 生成时需要采用自顶向下的方式，而对于频繁模式的挖掘使用的却是自底向上的方法进行处理。FP – Growth 算法使用递归方式生成条件 FP – tree，因此构造的 FP – tree 与条件 FP – tree 必须能够双向可遍历，这样系统就需要更多的存储空间来保存 FP – tree 和条件 FP – tree。因此，FP – Growth 算法在时间和空间效率上仍然有待提高，对于大数据的处理能力也不是太强。

本文提出的一种改进的 FP – Growth 算法，该算法改进了 FP – tree 结构，引入了一种前缀树结构 AFP – tree。对 AFP – tree 的挖掘采用深度优先的策略，不需要构造条件模式库，所以该算法能够显著提高挖掘效率。

(1) AFP – tree

前缀树 AFP – tree 记录的是事务数据库 DB 中的每个事务 Trans 的频繁项，FP – tree 中的项按支持度降序排列，在前缀树 AFP – tree 中，所有的频繁项组成一个偏序集，事务中的项按字典顺序排列。所以，两者的区别在于频繁项的顺序不同。AFP – tree 中的每个节点由 4 部分组成: 项目名称、频繁计数、节点指针和父节点指针。有 3 个频繁项的完全前缀树如图 3 所示，包含了 3 个频繁项所构成的所有模式。从根节点到其它节点表示每种模式，每个子树表示以该子树的根为前缀的所有模式。

从前缀树的结构可以看出，事务数据库的每条记录都存在于前缀树 AFP – tree 的某条路径中，前缀树 AFP – tree 的深度为数据库中事务的最大长度。前缀树 AFP – tree 的结构为我们提供了一个高效的挖掘策略: 如果某节点的计数小于最小支持度 minsup，也就是说该项不频繁，那么以该节点为根的整个子树中不存在频繁模式。因此，这种树结构可以显著提高挖掘的效率。当使用基于 FP – tree 的

FP – Growth 算法时，由于相同模式可能分布在不同的子树中，对单个子树进行挖掘时无法判断某个模式是否频繁，必须递归地构建条件 FP – tree 才能最终判断，效率不高。而挖掘前缀树 AFP – tree 就不需要构造条件模式树，采用深度优先的挖掘策略就可判断某个模式是否为频繁模式。

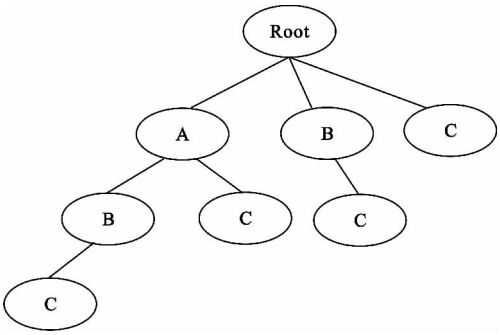

图 3 3 个频繁项的完全前缀树 AFP – tree
Fig. 3 Complete prefix tree of AFP – tree with three frequent items

(2) 基于 AFP – tree 的频繁模式挖掘算法

本算法采用由上而下的深度优先的策略进行挖掘，挖掘步骤分为以下几步:

1) 扫描事务数据库 DB 一遍，根据频繁项之间的偏序关系构造初始前缀树 AFP – tree，对各频繁项按照支持度计数的顺序形成头表。

2) 不断调整前缀树中的节点链和节点计数，从而在 AFP – tree 中采用由上而下的深度优先的挖掘策略来挖掘频繁模式。

算法的执行过程如下:

输入: 事务数据库 DB，最小支持度 min_sup。

输出: 频繁模式的完全集。

①生成 AFP – tree

对数据库项进行一遍扫描，获取各频繁项的集合 F 与其支持度的相关信息。按照支持度对集合 F 进行降序排序，获得频繁项表 L。

构建树的根部节点，其值记为“null”。然后对数据库中的每项事物 Trans 按以下方式处理: a) 将位于 Trans 中的各频繁项按照 L 当中的相同次序进行排序。假设排序以后的频繁项列表表示为 [p | P]，其中 p 是首个项目，P 是列表的其他部分; b) 调用方法 insert_tree([p | P], T), insert_tree([p | P], T) 的具体执行过程如下: 假如 T 中包含 1 个子结点 N，满足条件 N. item_name = p. item_name，则对 N 的计数器执行加 1 操作; 否则，按照偏序关系新建 1 个节点 N，使其计数为 1，通过父节点指针链接到它的父节点 T，并且使用节点链结构将其链接到 item_name 相同的节点上。如果 P 不是空集，则递归调用方法 insert_tree(P, N) 。

②对 AFP – tree 进行挖掘

对每个频繁项 α 执行以下过程:

```
Procedure AFP_Mine (α, AFP – tree)
for any α 的子树根节点 subroot do
if (α 的右兄弟节点中存在节点 ribroot，使得 ribroot. item = subroot. item) or (α 的右兄弟节点中不存在与节点 subroot 相同名称的节点，ribroot = null) then Combine (subroot, ribroot);
end for
  for anyα 的子节点 α_i do
      if α_i. count ≥ minsup then
将 α. α_i 和 α. α_i. count 加到频繁模式集;
          AFP_Mine (α. α_i, AFP – tree);
      end if
   end for
end
Procedure Combine (α,β)
  If β = = null then β = α;
else β. count = β. count + α. count;
      for anyα 的子节点 α_i do
设β 的子节点中与 α_i 名称相同的节点为 β_i，Combine (α_i,β_i);
      end for
end if
end
```

3 实验与分析

本文采用 XenServer 构建模拟云平台作为实验平台，并选取美国麻省理工学院林肯实验室公开提供的 DARPA 入侵检测评价计划中的数据集 KDD–Cup99[8] 进行实验，构建客户实例。整个 KDDCup99 数据集约有 490 万条数据记录，数据量很庞大。考虑到实验的可行性，从中选取 10% 的数据作为本实验的数据集。

分别选用较少、普通和很多攻击案例来模拟真实世界中的网络使用环境，可以得到 3 组不同的数据集合，各组数据集合中的训练集与测试集情况如表 1 所示。

使用本文提出的改进 FP – Growth 算法对表 1 中的 3 组训练集分别进行学习训练，进而提取特征模式，生成特征库，然后使用测试集分别对其进行测试，得到的实验结果如表 2 所示。

采用未改进的 FP – Growth 算法进行实验的结果见表 3。

比较表 2 和表 3 可以看出:

1) 本文中的各组训练时间分别比传统的 FP – Growth 算法降低 50.82%、52.40% 和 54.00%。因此,本文提出的 FP – Growth 改进算法可以在很大程度上提高系统的检测效率。

2) 对于各组数据,改进后的 FP – Growth 算法的检测效率比传统 FP – Growth 算法提高了 1.19% ~ 1.75%,误报率降低了 13.84% ~ 17.52%,漏报率降低 30.90% ~ 44.46%,检测性能得到了很大改善。

实验结果表明,本文提出的 FP – Growth 改进算法较传统算法在性能上有显著提高,非常适用于云环境下的安全防护。

表 1　数据集样本组成情况

Table 1 Composition of data sets

组	训练/测试	总数	正常	异常			
				DOS	R2L	U2R	Probing
一	训练	3 000	2 940	36	15	3	6
	测试	3 500	3 430	40	18	7	5
二	训练	3 500	3 220	160	75	20	25
	测试	4 000	3 680	185	80	18	37
三	训练	4 000	3 280	420	160	55	85
	测试	4 500	3 690	435	200	60	115

表 2　基于 FP – Growth 改进算法的实验结果

Table 2 Experimental results of the improved FP – Growth algorithm

组	检测率/(%)	误报率/(%)	漏报率/(%)	训练时间/s
一	95.71	4.08	14.28	13.12
二	94.85	4.46	13.13	15.96
三	93.38	5.83	10.25	18.85

表 3　改进前的 FP – Growth 算法实验结果

Table 3 Experimental results of the FP – Growth algorithm

组	检测率/(%)	误报率/(%)	漏报率/(%)	训练时间/s
一	94.69	4.90	25.71	26.68
二	93.37	5.30	21.88	33.53
三	91.47	6.94	15.80	40.98

4　结　语

本文提出一种基于关联规则挖掘的智能云防护技术,利用云平台中大量探针提取出的数据信息,在云端进行分析和处理,应用数据挖掘技术建立网络行为特征库。针对传统 FP – Growth 算法存在的时空效率不高,海量数据处理能力较弱的缺点,提出一种改进的 FP – Growth 算法,该算法改进了 FP – tree 结构,引入了一种前缀树结构。实验结果表明,改进算法能有效地提高数据挖掘的速度,增强检测能力,为网络安全维护提供了坚实的基础。

参考文献:

[1] HUANG Dijiang, ZHOU Zhibin, XU Le. Secure Data Processing Framework for Mobile Cloud Computing [C] //IEEE INFOCOM 2011 Workshop on Cloud Computing. Shanghai: IEEE, 2011: 614 – 618.

[2] SAKR Sherif, LIU Anna, BATISTA Daniel M. A Survey of Large Scale Data Management Approaches in Cloud Environments [J]. IEEE Communications Surveys & Tutorials, 2011, 13(03): 311 – 336.

[3] DIMITRIOS Zissis, DIMITRIOS Lekkas. Addressing Cloud Computing Security Issues [J]. Future Generation Computer Systems, 2012, 28(03): 583 – 592.

[4] HAN Jiawei, KAMBER Micheline. 数据挖掘——概念与技术 [M]. 北京: 机械工业出版社, 2001.
HAN Jiawei, KAMBER Micheline. Data Mining: Concepts and Techniques [M]. Peking: China Machine Press, 2001.

[5] 肖东荣,杨磊. 基于遗传算法的关联规则数据挖掘 [J]. 通信技术, 2010, 43(01): 205 – 207.
XIAO Dongrong, YANG Lei. Association Rule Data Mining Based on Genetic Algorithm [J]. Communications Technology, 2010, vol. 43(1): 205 – 207.

[6] RAMASWAMY Sridhar, RASTOGI Rajeev, SHIM Kyuseok. Efficient Algorithms for Mining Outliers from Large Data Sets [C] //Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, TX, USA: ACM, 2000: 427 – 438.

[7] SEQUEIRA Karlton, ZAKI Mohammed. ADMIT: Anomaly – based Data Mining for Intrusions [C] //Proceedings of the 8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada: ACM, 2002: 386 – 395.

[8] KDD Cup 1999 Data [EB/OL]. (1999 – 10 – 28) [2014 – 04 – 10]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

作者简介:

方忠进(1979 –),男,博士研究生,讲师,主要研究方向为网络信息安全、云安全;

FANG Zhong – jin (1979 –), male, Ph. D student, lecturer, majoring in network information security and cloud security.

夏志华(1983 –),男,博士,讲师,主要研究方向为信息安全、隐写分析;

XIA Zhi – hua (1983 –), male, Ph. D., lecturer, mainly engaged ininformation security and steganography.

周　舒(1984 –)女,硕士,助理研究员,主要研究方向为信息安全。

ZHOU Shu (1984 –), female, M. Sci., research associate, mainly working atinformation security.