基于关联规则的网络信息内容安全事件发现及其 Map-Reduce 实现

葛 琳* 季新生 江 涛
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要： 针对网络中信息内容安全事件的发现问题，该文提出一种基于关联规则的多维度用户行为特征关联分析法；对于存在的虚警问题，提出了基于邦弗朗尼校正的检验准则；为满足在海量数据中的应用需求，提出了一种 Map-Reduce 框架下的分布式幂集 Apriori 算法。实验结果表明，该文提出的方法及相应算法，并行运算能力强，在低虚警率和漏检率的情况下，具有较好的检测率，且运行时间短，收敛速度快。

关键词： 网络安全；关联规则；信息内容安全事件；Apriori 算法；邦弗朗尼校正；Map-Reduce
中图分类号： TP309 **文献标识码：** A **文章编号：** 1009-5896(2014)08-1831-07
DOI: 10.3724/SP.J.1146.2013.01272

Discovery of Network Information Content Security Incidents Based on Association Rules and Its Implementation in Map-Reduce

Ge Lin Ji Xin-Sheng Jiang Tao
(*National Digital Switching System Engineering and Technological Research Center*, *Zhengzhou* 450002, *China*)

Abstract: A multi-dimension association analysis method of user's behavioral characteristics based on association rules is proposed for the discovery of information content security incidents in network. The user's multi- dimension data which generate in communication can be mined. An inspection standard based on Bonferroni's correction is put forward to deal with the problem of false alarm. In order to meet the demand for the implementation of the method in a massive database, a distributed power set Apriori algorithm in Map-Reduce framework is proposed. Experimental results demonstrate that the proposed method and its corresponding algorithm have strong ability in parallel computing. The algorithm has a great detection rate in the case of low false alarm rate and missing detection rate. The running time is short and it can achieve a fast convergences rate.

Key words: Network security; Association rules; Information Content Security Incidents (ICSI); Apriori algorithm; Bonferroni's correction; Map-Reduce

1 引言

信息内容安全事件(Information Content Security Incidents, ICSI)是指利用信息网络发布、传播危害国家安全、社会稳定和公共利益的内容的安全事件[1]。随着网络中通信信息的海量增长，ICSI数目呈现上升趋势，对其中信息进行有效挖掘，实现信息内容安全事件的发现，已成为当前网络信息安全研究领域的重要组成部分和关注热点。

目前，信息内容安全领域的研究主要有：针对以传输特定信息为目的的信息渗透的检测技术研究[2]，针对网络信息内容安全的控制模型及评估框架的研究[3]，基于文本内容的事件分类技术[4]以及通过对多媒体内容的识别发现其中隐藏的安全事件[5]等。对于ICSI的发现技术尚属空白，本文则进行了这方面的尝试。现有的安全事件检测发现技术主要针对攻击或系统漏洞与缺陷：文献[6]提出了一种带权值的数据流频繁项挖掘算法，用于识别网络中的大流量对象；文献[7]采用基于近期最少使用算法(Least Recently Used, LRU)和空间代码布鲁姆过滤器(Space Code Bloom Filters, SCBF)的大象流提取方法，实现高速网络环境下大象流的及时准确提取，应用于对分布式拒绝服务(Distributed Denial of Service, DDoS)攻击的检测；文献[8]提出了一种网络设备协同联动模型，在僵尸网络的检测和追踪、DDoS攻击事件以及二者的关系分析中进行验证；文献[9]提出了一种协同方法，通过减少入侵防御/检测系统(Intrusion Prevention System/Intrusion Detection System, IPS/IDS)中的错误和低重要性告警，提高网络管理员对安全事件的聚焦能力；文献[10]提出了一种基于规则监测(Order-Based

2013-08-22 收到，2014-03-05 改回
国家 863 计划项目(2011AA010605)和国家科技重大专项(2012ZX03006002-010)资助课题
*通信作者：葛琳 lingesnow@126.com

Monitoring, OBM)方法，既可实现基于代理的(需要在监控网络中安装软件)又可实现无代理(通过内置系统协议如SNMP和Windows管理规范等)的计算机网络的安全监控。

事实上，相较于攻击类事件，ICSI以传播不良信息内容为目的，强度小、信息繁杂且具有隐蔽性，不会破坏网络中硬件设备导致故障和瘫痪等，也不存在类似IDS日志和系统漏洞等的告警信息。因此，对该类事件的发现难度较大。网络通信数据库中隐藏着诸多有用信息，关联规则可以发现数据库项集之间的关联关系。ICSI由用户的通信行为生成，通过对此类数据的分析，可以得到用户的呼叫行为(通信时长、通信类型)、位置移动(当前位置、涉及基站个数)、通信关系(与其他网络用户的通联)和通信内容(是否含有关键词)等特征，提取其中具有一定关联规则的多维数据形成判别事件的4个要素(时间、地点、人物和内容)，从而实现对ICSI的发现。Apriori算法[11]是最为经典的关联规则挖掘算法，Map-Reduce是现今较为流行的并行编程框架，用于大规模数据集的并行计算，其开源实现是Hadoop平台[12,13]。近年来，针对Map-Reduce的研究不断深入，主要围绕在改进Map-Reduce架构以适应处理某些特定应用，和将某些现有问题在Map-Reduce框架下处理。但是，对于基于Map-Reduce框架实现Apriori算法的相关研究还较为初步：文献[14]对如何实现Map-Reduce-Apriori算法进行了论述，并未对算法提出优化；文献[15]主要对Map-Reduce的框架结构进行了优化，采用Apriori算法对其进行性能的验证；文献[16,17]分别提出一种水平划分数据集和迭代式的Apriori优化算法并应用于Map-Reduce。现有方法仅仅是对Apriori算法和Map-Reduce的简单组合，没有针对具体的需求提出有建设性的应用方案。

据此，本文给出了一种基于关联规则的ICSI发现方法及其在Map-Reduce下的实现算法。实验结果表明，该算法具有良好的并行运算能力，在较低虚警率及漏检率的情况下，具有较高的检测率。本文的主要贡献如下：(1)提出一种基于关联规则的多维度通信信息关联的ICSI发现方法；(2)提出一种基于邦弗朗尼校正原理的要素数据集检验准则；(3)提出一种Map-Reduce框架下的分布式幂集Apriori算法；(4)通过大量实验和实际数据集测试，验证了本文提出方法的可行性和算法的有效性。

2 网络信息内容安全事件发现技术

网络多维通信信息中，隐含了各类安全事件发生的要素，如何提取出其中的有用信息，是发现ICSI的前提和基础。本节所分析数据的来源为VAST 2008中的Cell Phone Social Network数据集，包含400人、10天的 9834条通信记录。

2.1 多维通信信息的分析与关联

2.1.1 呼叫行为特征 图 1 描述了用户通信次数的分布。从图中可以看出，大部分用户的通话次数保持在 23~31 次，其中 27 次的用户数目最多。然而，有 4 个用户的通话次数大于 38 次，约占通信总人数的 1.00%。这种极少数用户的通信行为，通常表明了该地址/号码具有特殊用途，比如，作为商业号码或者 ICSI 的发送源。虽然均具有高通信次数，但前者属于合法通信行为，而后者则是需要观测的对象，需结合通信内容作出判别。

图 2 描述了用户通信时长的分布，其分布特征与图 1 类似。对于大部分用户，通信时长维持在 80~120 s 左右，其中 110 s 左右的人数最多，占通信总人数的 37.50%，只有少部分用户的通话时长短于 10 s 或者高于 220 s。具有超长通信时长的地址/号码，表现出了非常规性，同样，超短通话累积次数较多的地址/号码也往往可能具有某种特殊性。

2.1.2 位置移动特征 用户通信行为所涉及的基站信息能够从一定程度上反映用户的移动规律。图 3 描述了用户所涉及的基站数目分布，55.00%的用户涉及到的基站个数为 2，这也从侧面表明了用户活动一般具有地域局限性。那些涉及基站个数较多的用户，属于极少数。图 3 中，在 5 个基站下通信过的用户数目有 1 人，仅占通信总人数的 0.25%，代表了此用户的特殊性，可以将其作为观测对象。

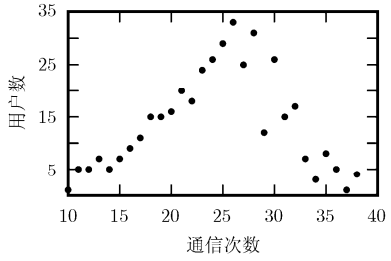


图 1 通信次数分布图

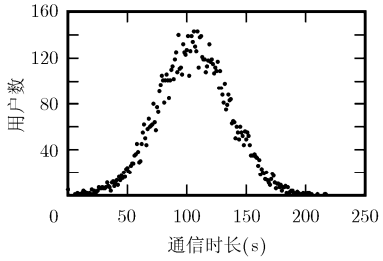


图 2 通信时长分布图

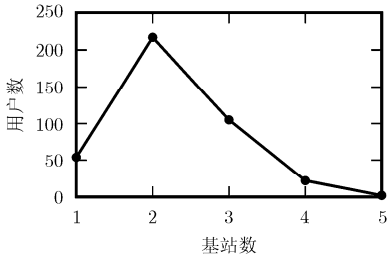


图 3 用户移动性分布图

2.1.3 通信关系特征　图 4 中表示的是用户的通联关系度。大多数用户通常具有较少的通信联络对象，即其关联度在 5 以下。其中，关联度为 2 的用户数目达 350 人，占总通信总人数的 87.50%。关联度超过 15 的用户只有 1 人，约占总通信人数的 0.25%。由此可以看出，那些通信对象较多的地址/号码通常可能具有特殊用途，是需要进一步关注的对象。

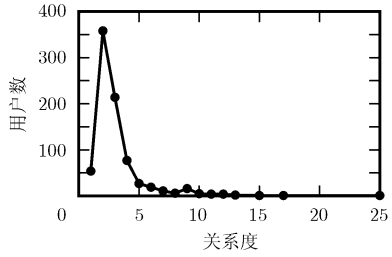


图 4 用户通联关系度分布图

时间、地点、人物和内容是能够清楚描述一件事物的 4 要素。其中，时间是预先设定的研究区间，地点由用户的位置及其移动性体现，人物反映用户的通联关系，内容则由通信行为和通信内容决定。将用户的呼叫行为和通信关系进行关联，可以划分出重点观测用户的地址/号码；对此类用户通信的具体内容进行进一步的分析；同时，针对重点通信用户的位置移动特征进行跟踪；综合得到的各要素信息，给出是否发现网络中存在 ICSI 的判定结果，如图 5 所示。

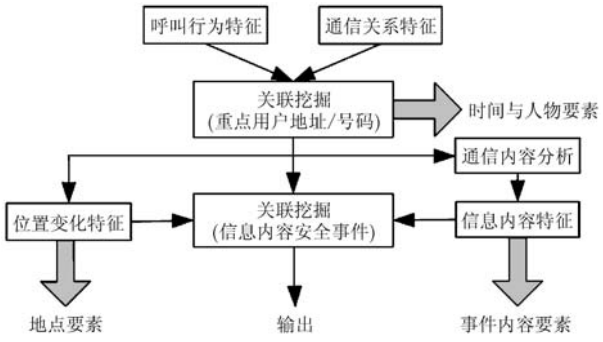


图 5 各要素的关联分析

2.2 Apriori 算法在信息内容安全事件特征分析中的应用

对事件各要素信息间的关联挖掘是实现 ICSI 发现的关键，而寻找频繁项集是关联规则挖掘的核心问题。本文采用经典的 Apriori 算法作为关联规则挖掘算法。

定义 1　某时间窗 T_i 内，$f_i \le \mathrm{ID}_i, T_i, \mathrm{Act}_i, \mathrm{Rel}_i, \mathrm{Loc}_i, \mathrm{Con}_i >$ 为用户 ID_i 的事实表，代表在该时间窗内此用户的所有通信记录。其中，Act_i , Rel_i , Loc_i 和 Con_i 为该事实表中的关键属性，分别代表用户 ID_i 的呼叫行为、通信关系、位置移动和通信内容。

定义 2　$\mathrm{Act}_i =< \mathrm{Long}_i, \mathrm{Short}_i, \mathrm{Type}_i >$, $\mathrm{Rel}_i =< \mathrm{ID}_j, \mathrm{Position}_j >$, $\mathrm{Loc}_i =< \mathrm{Degree}_i, \mathrm{Population}_i, \mathrm{Position}_i, \mathrm{Numbers}_i >$ 和 $\mathrm{Con}_i =< \mathrm{Keywords}_i >$ 分别表示与 ID_i 事实表关联的维度表。其中，Long_i 和 Short_i 分别表示是否含有超长和超短通话，设置值为 0 或 1; Type_i 表示通信的类型(语音、视频、短信和彩信等)；ID_j 表示被叫用户 j 的地址/号码；$\mathrm{Position}_i$ 和 $\mathrm{Position}_j$ 分别表示主叫用户 i 、被叫用户 j 的位置信息；$\mathrm{Numbers}_i$ 表示用户所涉及的基站数目；$\mathrm{Population}_i$ 表示该基站下的用户数；Degree_i 表示基站的通联度；$\mathrm{Keywords}_i$ 代表通信内容关键词。

定义 3　设 D 是网络中各项通信信息的集合，与 ICSI 相关的数据 I 是数据库事务的集合，其中每个事务 H 是项的集合，使得 $H \subseteq D$。设 A 是一个项集，事务 H 包含 A，当且仅当 $A \subseteq H$。关联规则是形如 $A \Rightarrow B$ 的蕴涵式，其中 $A \subset D$, $B \subset D$，并且 $A \cap B = \varnothing$。规则 $A \Rightarrow B$ 在事务集 I 中具有支持度 s，表示为：$\sup(A \Rightarrow B) = P(A \cup B)$；规则 $A \Rightarrow B$ 在事务集 I 中具有置信度 c ,表示为：$\mathrm{conf}(A \Rightarrow B) = P(B|A)$。

定义 4　同时满足最小支持度阈值 min_sup 和最小置信度阈值 min_conf 的规则为强规则。

事实表与各维度表的关系如图 6 所示。本文将网络中产生的所有用户通信记录作为数据集 D。首先，进行数据预处理，生成各 ID_i 的事实表和维度表；其次，对 Act_i 和 Rel_i 两个维度表，按照 $\mathrm{min\_sup}_1 = a$ 进行关联挖掘，找出所有的频繁项集，根据 $\mathrm{min\_conf}_1 = b$ 产生重点用户的强关联规则 1; 再次，将规则 1 作为约束条件，分别反馈给 Loc_i 和 Con_i 维度表，对表项进行过滤处理；最后，将关联规则 1 与约减后的 Loc_i 和 Con_i 进行第 2 次关联挖掘，找出所有的频繁项集，生成事件的强关联规则 2，按照 $\mathrm{min\_sup}_2 = \eta$ 和 $\mathrm{min\_conf}_2 = \gamma$ 输出最终结果。

2.3 邦弗朗尼校正与信息内容安全事件发现

文献[13]提到：随着数据规模的增长，某类特定事件会出现的数目也随之上升。然而，诸多的随机数据往往都会具有一些非寻常特征，这些特征虽然看上去重要，但实际上并不重要，据此类数据推断出的事件出现属于“Bogus”。当数据量增长到一定规模以后，可以从小量数据中挖掘出有效信息的算法并不一定适用于大数据。由此可以看出，提取到

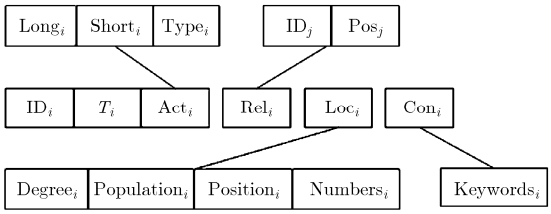


图 6 事实表与维度表的关系

事件各类要素数据并非完全可信、可用，如果对其直接应用极有可能产生虚警事件。那么，如何避免将随机出现看成真正的出现，统计学中的邦弗朗尼校正(Bonferroni's correction)给出了一个在统计上可行的方法，来规避在搜索数据时出现的“Bogus”正响应。根据邦弗朗尼校正的思想，本文提出 ICSI 要素信息的检验准则。

定理 在网络通信数据集 D 中，通过关联规则得到由 N 项组成的频繁项集 Z，即 $Z=\{\boldsymbol{m}_1,\boldsymbol{m}_2,\cdots,\boldsymbol{m}_i,\cdots,\boldsymbol{m}_N\}$（$\boldsymbol{m}_i$ 为网络中安全事件要素组成的多维向量)。为满足判定频繁项集 Z 为安全事件来源信息的概率低于显著水平 α，则需 $P_i\le\alpha/N$（P_i 为第 $\boldsymbol{m}_i$ 项对此事件成立的支持度)。

证明 设 P 为频繁项集 Z 犯第 1 类错误的概率，n_j 为 N 项中不为空且假设成立的子集数目。根据布尔不等式得

$$\mathrm{P}\left\{\bigcup_{i=1}^{N}\left(P_i\le\frac{\alpha}{N}\right)\right\}\le\sum_{i=1}^{N}\left[\mathrm{P}\left(P_i\le\frac{\alpha}{N}\right)\right]\le n_j\times\frac{\alpha}{N}\le N\times\frac{\alpha}{N}=\alpha$$

证毕

过程 通过关联规则得到的频繁项集 Z 中，共有 N 项通信信息。根据邦弗朗尼原理，并非每项消息都是可用的。将 N 项消息按照 P_i 值由小到大排序，每次取出最小的 P_i 值进行判定，即

若 $P_1\le\alpha/N$，则判为显著，否则判为不显著；对第 2 项进行判别，若 $P_2\le\alpha/(N-1)$，则判为显著，否则判为不显著；以此类推，对于第 j 项，若 $P_j>\alpha/(N-j)$，则判为不显著，且后续 P_i 检验都不显著。根据邦弗朗尼原理，对提取出的频繁项集再进行约减，得出判为显著的项集，才构成 ICSI 发现的数据集。该方法可以有效提高 ICSI 的判别准确度，降低虚警率。其中 P_i 值的求取，将在下面进行介绍。

2.4 网络信息内容安全事件发现流程

ICSI 的发生由用户的异常通信行为引起，本文将在时间窗 T_i 内频繁发生且不被网络正常行为所覆盖的模式作为事件模式；将在正常情况下，数据集中频繁出现的正常通信行为作为正常模式。正常模式与事件模式中，均含有各自的频繁项集，将两类项集进行相似度匹配，求得的匹配度即为上文中提到的项对事件的支持度 P_i，发现流程如图 7 所示。

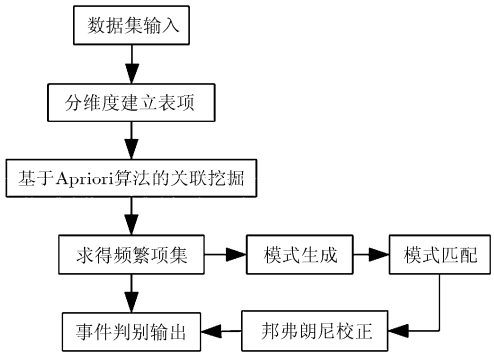


图 7 信息内容安全事件发现流程示意

定义 5 若频繁项集 X 的所有超集都是非频繁项集，则称 X 为最大频繁项集(Maximum Frequent Item Set, MFIS)，将所有最大频繁项集组成的集合称为最大频繁集(Maximum Frequent Set, MFS)。

定义 6 设 A 和 B 是逻辑同构的两个 N 项数据集合，各自的最大频繁集分别为：$\mathrm{MFSA}=\{(A_1,C_{\mathrm{A}1}),(A_2,C_{\mathrm{A}2}),\cdots,(A_i,C_{\mathrm{A}i}),\cdots,(A_N,C_{\mathrm{A}N})\}$，$\mathrm{MFSB}=\{(B_1,C_{\mathrm{B}1}),(B_2,C_{\mathrm{B}2}),\cdots,(B_j,C_{\mathrm{B}j}),\cdots,(B_N,C_{\mathrm{B}N})\}$。其中，$(A_i,C_{\mathrm{A}i})$ 为 A 的第 i 项最大频繁项集及其支持度，$(B_j,C_{\mathrm{B}j})$ 为 B 的第 j 项最大频繁项集及其支持度。

本文根据定义 5 和定义 6 实现模式生成，模式匹配部分采用文献[18]中，对逻辑同构的两个数据集相似性的计算方法，对 A 和 B 的相似性，即 $\mathrm{Sim}(A,B)$ 进行计算，并将其值赋予 P_i，即 $\mathrm{Sim}(A,B)\to P_i$。

3 Map-Reduce 架构下的 DPS-Apriori 算法

本文提出了一种应用于 Map-Reduce 架构下的分布式幂集 Apriori 算法(Distributed Power Set Apriori, DPS-Apriori)，提高频繁项集挖掘的效率。通过上文分析，将网络的通信记录数据集 D，按维度分解为 4 个不相交的子集交付节点处理，每个节点含有 map 函数映射键值对 $<\mathrm{key},\mathrm{value}>$，再将其传至 reduce 函数进行组合，根据预设的关联规则，求出频繁项集，如图 8 所示。图中，各个维度子集的候选集及频繁项集 L_{k0}, L_k 均采用 Power Set Apriori 算法进行搜索，这种分布式(distributed)的算法结构更适用于 Map-Reduce 架构。该算法运算简单，运行速度较快，实现了通过 2 次扫描数据库、1 次剪枝，就能挖掘出所有频繁项的功能，较适用于具有多条记录，但只有较少的项的数据库的频繁

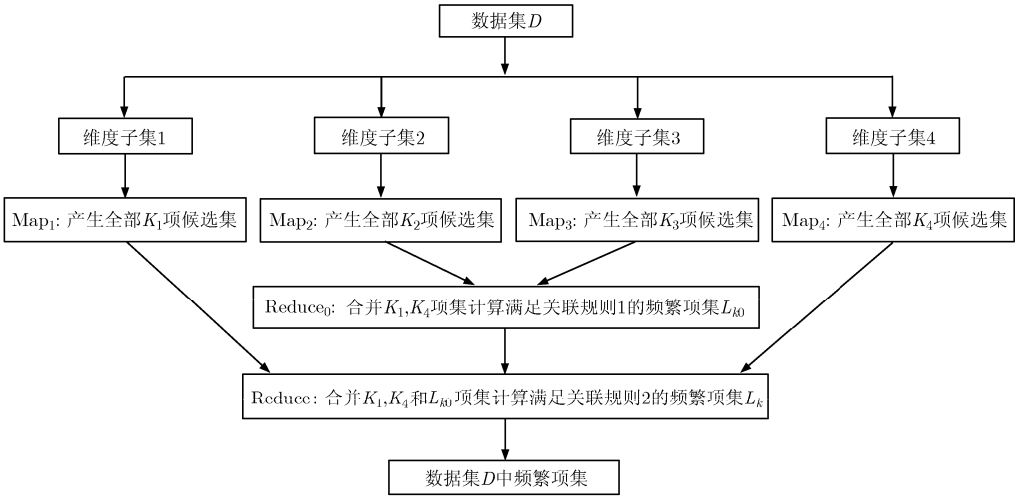


图 8　Map-Reduce-DPS-Apriori 结构

项搜索。算法的步骤为：

步骤 1　对数据集 D 进行第 1 遍扫描，获得每个候选项的计数，从而获得频繁 1 项集；

步骤 2　根据最小支持度对频繁 1 项集进行筛选；

步骤 3　幂集法生成所有筛选后的 1 项集的子项作为候选项；

步骤 4　对数据集 D 进行第 2 遍扫描，获得每个候选项的计数；

步骤 5　根据性质若 K 为频繁项集，则 K 的任何非空子集都是频繁项集，进行剪枝；

步骤 6　得出 MFIS。

4　实验评估

4.1 实验设计

本文实验的数据集来源为依托项目实验的电信网络数据，该数据集由网络中用户通信产生，对用户隐私进行了隐匿和屏蔽处理。参照VAST 2008中Cell Phone Social Network数据集的格式，测试数据集包括前文分析过的4个维度，即 Act_i, Rel_i, Loc_i 和 Con_i ； 共 10 项， 即 Long_i, Short_i, Type_i, ID_j, $\mathrm{Position}_i$, $\mathrm{Position}_j$, $\mathrm{Numbers}_i$, $\mathrm{Population}_i$, Degree_i 和 $\mathrm{Keywords}_i$。为测试本文ICSI发现方法和算法的可行性和有效性，实验中将通过特定设置(如，关键词设定等)产生ICSI。测试数据集的总数据量为10 G，共约8000万条通信记录，可判为ICSI的通信记录条数约为80万条。实验中，主要的对比算法为本文提出的Map-Reduce-DPS-Apriori算法与现有的Map-Reduce-Apriori算法。为更有力验证本文算法的有效性，通过前期测试，分别将两个算法的性能调整至最优。根据确保算法综合性能最佳的选取原则，本文将Map-Reduce-DPS-Apriori算法的支持度取值定为0.40，现有Map-Reduce-Apriori的支持度取值定为0.63。

实验所采用的主要测试指标分为两类，一类为测试算法并行计算能力的可扩展性(scaleup)、规模增长性(sizeup)和加速比(speedup)；另一类为测试算法检测性能的漏检率和虚警率。

漏检率=[(ICSI 总数目 – 正确检测到的 ICSI 数目)/ ICSI 总数目]×100%；

虚警率=[(检测到的 ICSI 数目 – 正确检测到的 ICSI 数目)/检测到的 ICSI 数目]×100%。

4.2 实验结果分析

4.2.1 算法的并行计算能力评估　图 9 表示的是算法的可扩展性能力。从图中可以看出，在同时增加数据规模和节点数目的情况下，本文算法的 Scaleup 值在理想值 1 附近。随着节点数目的增加，系统内部的通信开销额外增加，随着节点数目的增加，通信开销越来越大，Scaleup 值呈现下降趋势，不可能达到理想的 1 值。但本文算法的 DPS-Apriori 算法计算速度快，减少了数据规模增加为节点带来的运算负荷，也从某种程度上抵消了部分由于增加节点所带来的系统通信开销。因此， Scaleup 值整体波动幅度较小，体现了算法具有较好的可扩展能力。

图 10 表示了算法的规模增长性能力，将 1 G 数据作为初始的基准数据。从图中可以看出，在保持节点数目不变，增加数据集规模的情况下，本文算法的 Sizeup 值的增长率呈现为凸函数。随着数据规模的增加，算法的运行时间曲线整体为上升趋势，但随着数据增加而趋于平缓。分析原因为数据规模的增大相应增加了候选项集，占用了大量通讯和处理的开销，势必会引起处理时间的增长。但是本文

算法中的关联分析思想，使得数据集规模增加到一定程度之后，将增加的基准数据作为增量数据，即在前面已有的数据分析基础上再进行关联，减少了系统的运行时间。

图 11 为算法的加速比。将 1 G, 2 G, 3 G, 4 G, 5 G 和 6 G 数据分别作为 1~6 倍的基准数据。图 11 中，固定基准数据集的规模，增加节点的数目，各个基准数据集下的 Speedup 值围绕在直线 $y=x$ 附近，拟合程度较高。随着节点数目的增加，算法的运行时间增长，Speedup 值增大，同时，各个折线与直线 $y=x$ 接近重合，说明本文算法的并行化效果较好。从图中可以看出，随着基准数据量的增加，拟合程度愈来愈好，6 倍基准数据的 Speedup 值几乎与 $y=x$ 重合，说明本文算法在大规模数据的应用中性能较佳。

4.2.2 算法随维度和数据量的变化 图 12 表示的为两种算法随数据维度的变化情况，此处的数据维度变化指项集即特征类别的增加。从图中可以看出，随着维度的增加，本文算法的各个指标均优于现有算法，漏检率低于现有算法，虚警率则较现有算法有较大幅度的改进，且随着维度在 7 以上增加，虚警率基本维持在一个稳定的较低值，变化极小。

图 13 表示的为算法随数据量的变化，从图中可以看出，随着数据量的增加，现有算法的虚警率呈现上升趋势，而本文算法的虚警率始终变化水平较低。本文算法具有该性能的主要原因为基于邦弗朗尼原理建立的检验准则，有效滤除了计算结果中的与 ICSI 无关的频繁项集。降低虚警率可以有效提升算法的可靠性，减少网络安全管理的工作量，在实际应用中具有重要意义。因此，本文算法在处理高维项集和多数据库关联分析的问题上具有显著优势，对 ICSI 的发现具有有效性和可靠性。

4.2.3 算法随节点数量的变化 图 14 表示了两种算法在数据量为 6 G 情况下，随 Map 节点数的变化，本文算法的计算时间大大小于现有算法，且在节点数 4 之后基本处于稳定状态，两种算法的运行时间随着节点数的增加逐渐接近。图 14 中的运行时间主要由两部分构成，一是 Map-Reduce 系统内部的调度和通信时间，二是各个节点对数据集进行的计算时间。当节点数目较少时，系统消耗时间也较少，节点的计算时间占据主导，而本文提出的 Map-Reduce-DPS-Apriori 算法的优势主要体现在对数据集的计算时间短，因此此时本文算法与现有算法的运行时间差距明显；当节点数目逐渐增多时，系统消耗时间随之增大占据主导，节点中运算速度快的优势不再显著，使得两种算法的运行时间差距变小。由此可以看出：(1) 本文提出的 Map-Reduce-DPS-Apriori 算法运行时间短，对事件检测的时效性强；(2)本文算法在节点数目较少时即可达到整体性能的最优，即在较短的运行时间内具有较高的检测率，节约了系统运行成本，更加绿色、节能。

以上实验结果表明，本文提出的多维度关联 ICSI 发现方法及其实现算法(DPS-Apriori)具有良好的并行计算能力，在保持较高检测率的情况下，具有较低的虚警率和漏检率，在 ICSI 的应用中具有有效性和可靠性，且其运行时间短、收敛速度快。

5 结束语

为解决在当前海量网络通信数据中挖掘有用信息，实现信息内容安全事件发现的问题，本文给出了一种从用户行为分析作为切入点的解决方法。重点研究了网络中通信用户的行为特征，并将其数据信息分维度讨论，针对各维度中存在的非常规用户行为，提出了基于分布式幂集 Apriori 算法，采用了分别挖掘各维度中的频繁项集，再进行关联分析的方法。同时，为了降低虚警事件对最终判决结果的影响，提出了基于邦弗朗尼校正原理的检验准则，进一步对计算结果中的频繁项进行筛选。最后，在 Map-Reduce 框架下验证了本文所提出的事件发现方法及相应算法。实验结果表明，本文所提出的信

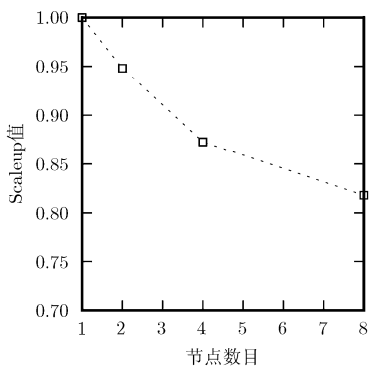


图 9 Scaleup 性能指标

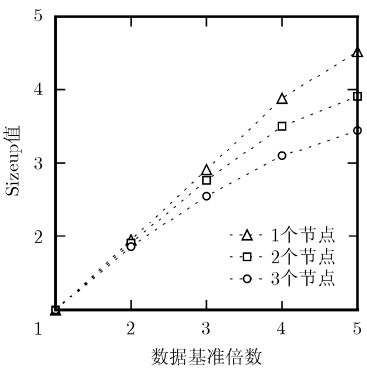


图 10 Sizeup 性能指标

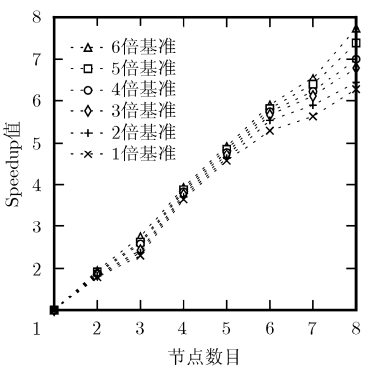
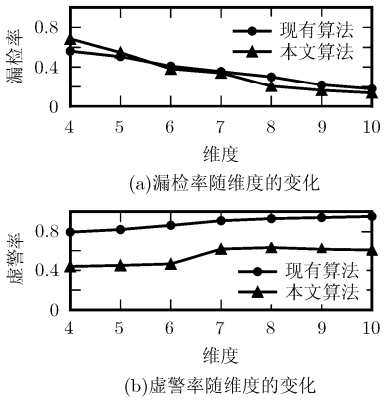


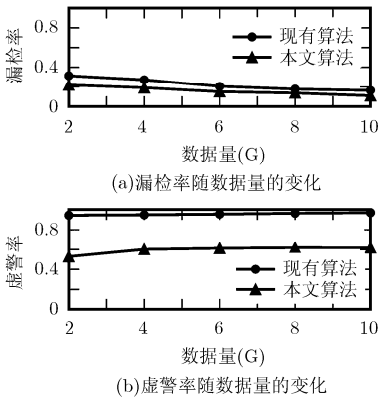
图 11 Speedup 性能指标



(a)漏检率随维度的变化

(b)虚警率随维度的变化

图 12 算法随数据维度的变化



(a)漏检率随数据量的变化

(b)虚警率随数据量的变化

图 13 算法随数据量的变化

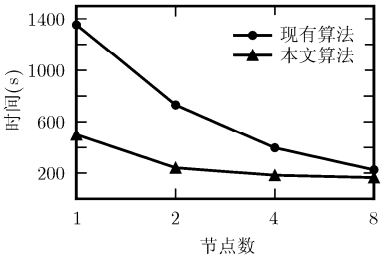


图 14 算法随节点数的变化

息内容安全事件的发现方法行之有效，相应算法性能明显优于现有算法。

参 考 文 献

[1] 国家标准化管理委员会. GB/Z 20986-2007 信息安全事件分类分级指南[S]. 北京: 国家标准化管理委员会, 2007.

[2] 陈训逊, 方滨兴, 胡铭曾, 等. 一个网络信息内容安全的新领域—网络信息渗透检测技术[J]. 通信学报, 2004, 25(7): 185–191.

[3] Fang Bin-xing, Guo Yun-chuan, and Zhou Yuan. Information content security on the Internet: the control model and its evaluation[J]. *SCIENCE CHINA*: *Information Sciences*, 2010, 53(1): 30–49.

[4] 万源. 基于语义统计分析的网络舆情挖掘技术研究[D]. [博士论文], 武汉理工大学, 2012.

[5] Barroso N, Lopez de Ipina K L, Ezeiza A, *et al.*. An ontology-driven semantic speech recognition system for security tasks[C]. Proceedings of IEEE International Carnahan Conference on Security Technology, Barcelona, Spain, 2011: 1–6.

[6] 张玉, 方滨兴, 张永铮. 高速网络监控中大流量对象的识别[J]. 中国科学: 信息科学, 2010, 40(2): 340–355.

[7] 谢冬青, 周再红, 骆嘉伟. 基于 LRU 和 SCBF 的大象流提取及其在 DDoS 防御中的应用[J]. 计算机研究与发展, 2011, 48(8): 1517–1523.

[8] 臧天宁, 云晓春, 张永铮, 等. 网络设备协同联动模型[J]. 计算机学报, 2011, 34(2): 216–228.

[9] Aguirre I and Alonso S. Improving the automation of security information management: a collaborative approach[J]. *IEEE Security & Privacy*, 2012, 10(1): 55–59.

[10] Kufel L. Security event monitoring in a distributed systems environment[J]. *IEEE Security & Privacy*, 2013, 11(1): 36–43.

[11] Agrawal R, Faloutsos C, and Swami A. Efficient similarity search in sequence databases[C]. Proceedings of the Fourth International Conference on Foundations of Data Organization and Algorithms, Chicago, 1993: 69–84.

[12] Dean J and Ghemawat S. MapReduce: simplied data processing on large clusters[J]. *Communication of the ACM*, 2008, 51(1): 107–113.

[13] Rajaraman Anand and Ullman Jeffery David. Mining of Massive Datasets[M]. London: Cambridge University Press, 2011: 67–69.

[14] Cryans J D, Ratté S, and Champagne R. Adaptation of Apriori to MapReduce to build a warehouse of relations between named entities across the Web[C]. Proceedings of Second International Conference on Advances in Databases, Knowledge, and Data Applications, Montreal, 2010: 185–189.

[15] Li Ning, Zeng Li, He Qing, *et al.*. Parallel implementation of Apriori algorithm based on MapReduce[C]. Proceedings of 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Kyoto, 2012: 236–241.

[16] Li Ling-juan and Zhang Min. The strategy of mining association rule based on cloud computing[C]. Proceedings of International Conference on Business Computing and Global Information, Shanghai, 2011: 475–478.

[17] Yang Xin-yue, Liu Zhen, and Fu Yan. MapReduce as a programming model for association rules algorithm on Hadoop[C]. Proceedings of 3rd International Conference on Information Sciences and Interaction Sciences (ICIS), Chengdu, 2010: 99–102.

[18] 毛伊敏, 杨路明, 陈志刚, 等. 基于数据流挖掘技术的入侵检测模型与算法[J]. 中南大学学报(自然科学版), 2011, 42(9): 2720–2728.

葛　琳：　女，1978年生，博士生，研究方向为网络信息安全.

季新生：　男，1969年生，教授，博士生导师，研究方向为网络安全.

江　涛：　男，1974年生，讲师，研究方向为移动互联网安全.