事件关联在证据链构造中的研究

刘 栋[1,2],宁玉富[1,2]
(1. 山东青年政治学院,山东 济南 250103;
2. 山东省高校信息安全与智能控制重点实验室,山东 济南 250103)

摘 要:在电子取证工作中,往往忽略对电子证据信息的预处理,从而导致电子证据冗余较大,计算分析较复杂。为解决计算机取证中存在电子证据形式化表示的困难以及数据缺失的问题,在对事件关联技术进行研究和深入分析的基础上,利用贝叶斯网络理论,提出一种基于事件关联的证据链构造方法。该方法考虑事件之间的相互影响以及序列关系,分析缺失数据的因果关系,拟合完整证据链,实现了形式化表示电子证据,并降低了证据分析的数据冗余,从而有针对性地进行数据处理和证据分析,完善了取证体制。通过实验结果分析得出,该方法实现了证据的形式化表示,减少了证据分析的数据量。

关键词:计算机取证;事件关联;贝叶斯网络;证据链;电子证据
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2016)12-0107-04
doi:10.3969/j.issn.1673-629X.2016.12.024

Research on Event Correlation in Construction of Evidence Chain

LIU Dong[1,2],NING Yu-fu[1,2]
(1. Shandong Youth University of Political Science,Jinan 250103,China;
2. Key Laboratory of Information Security and Intelligent Control of Shandong Universities,
Jinan 250103,China)

Abstract:The electronic evidence data preprocessing is easily neglected in electronic forensics work,leading to heavy redundancy for electronic evidence and complex calculation. Since the electronic evidence is difficult to represent formalized,and there exist missing data. A method for constructing electronic evidence chain is proposed on the basis of the study and analysis of event correlation and Bayesian network. Considering the interaction between evidence events and sequence relationship,it can be analysis of causal relationship of the events to deal with the missing data. It realizes the electronic evidence represented and reduces the data redundancy of evidence analysis,thus consummating the evidence collection system and making the data process and evidence analysis be more target-oriented. The experimental results show that the method realizes the representation of evidence and reduces the computation.

Key words:computer forensics;event correlation;Bayesian network;evidence chain;electronic evidence

0 引 言

计算机取证是解决争议和打击计算机犯罪的重要手段,专门研究如何按照符合法律规范的方式收集、处理计算机犯罪证据,是实现信息安全保障的一个重要方面,在保持社会稳定和维护法律秩序方面具有重要作用[1-2]。近年来,随着计算机取证不断在实用性和有效性方面的深入研究,在电子证据的获取、分析、表示等方面取得了许多经验和进展。文献[3]用手工定义的入侵事件间概率相似度和极小匹配规则来构建入侵事件关联专家系统,但难以直接获取所需知识;文献[4]将前提和目的吻合的入侵事件关联形成入侵者攻击轨迹,但主观性较强;文献[5]提出了运用关联算法分析事件间的关系;文献[6]提出一种基于计划的事件关联模型,但分析数据量较大,未考虑数据缺失的情况。

文中在总结前人研究的基础上,提出一种基于事件关联的证据链构造方法,用于证据分析,便于形式化表示证据,完善了取证体制。

收稿日期:2016-02-14 **修回日期**:2016-05-19 **网络出版时间**:2016-11-21
基金项目:国家自然科学基金资助项目(60873247)
作者简介:刘 栋(1987-),男,硕士研究生,助理实验师,研究方向为数据挖掘;宁玉富,教授,博士,硕士生导师,研究方向为不确定理论。

1 相关概念

先前的证据分析大都是对事物间的统计关系的挖掘,未涉及底层因果结构以及电子证据间的相关性,而电子证据的表现方式多样,获取的证据不一定就是完整数据,对于证据的形式化出示一直存在困难[7-8]。为了支持司法分析,有必要进行证据链的挖掘与构造。证据链的构造具体为事件的关联分析,其基本内容主要包括将每一子事件以时间序列重新定义,进而把不同的事件联系起来,挖掘深层次、有价值的模式。事件关联分析的目的是进行数据预处理,主要包括信息计数、数据浓缩、信息抑制和事件概括[9]。

在计算机取证领域,很多专家学者已经做了大量深入研究,文中从理论层面研究事件关联分析在证据链构造中的应用。

定义 1 事件[10](Event):是指在计算机系统中由某项活动产生的记录。

定义 2 原子事件(Atom Event,AE):描述用户一次具体的请求(e),比如仅单击某个按钮。

2 证据链的构造

证据链的构造问题可以描述成:将证据链 E 分割成原子事件集合 $E=\{e_1,e_2,\cdots,e_i,\cdots,e_n\}$,按照某种规则决定原子事件 e_i 在原始证据链中的顺序。为决定原子事件恰当的顺序,需识别证据中邻近碎片对。用 $p(i,j)$ 表示原子事件 e_j 接在 e_i 后面的可能性,称 $p(i,j)$ 为权重。在所有可能原子事件序列中,接近正确的序列顺序是权重值和为最大(或最小)的序列。设 W 为原子事件某序列 π 的权重值,则 $W=\sum_{i=0}^{n=1}P(\pi(i),\pi(i+1))$。

2.1 事件的关联分析

对于事件之间的关系,文中列出了事件之间常见的关联关系类型,具体如下:

(1)压缩(compression):去除冗余的过程,计算多个事件的关联度,将多事件归纳为单事件,形式为:$[e,e,e,\cdots]\Rightarrow e$。

(2)过滤(filtering):假定源事件集 e 的属性诸多 $M(e)$ 不属于目的事件集,则过滤掉源事件集 e 中的该类事件,形式为:$[e,M(e)\notin H]\Rightarrow\varphi$。

(3)压制(suppression):将事件进行优先级排列,形式为:$[e,E]\Rightarrow E$。

(4)计数(count):重复事件的计数归纳过程,尽可能减少重复计算,形式为:$[n\times e]\Rightarrow E$。

(5)时序关系(temporal relation):根据依赖函数,计算事件之间的时间序列模式,形式为:$[|e_1-e_2|<T]\Rightarrow e_3$。

2.2 证据链构造方法

根据协同取证的原理,构造证据事件的序列模式关键在于子事件内部序参量之间的协同作用[11]。证据事件关联系统可以定义为一种自组织结构,在有赖于事件之间的关联作用的有序组织中相继发生的事件之间通过相互作用形成,事件关联度用来描述事件之间相互关联相互影响的程度。

(1)作用函数。

α_{ij},β_{ij} 是系统稳定临界点上序参量的上、下限值。对原子事件有序的作用系数 u_{ij} 可表示为:

$$u_{ij}=\frac{(X_{ij}-\beta_{ij})/(\alpha_{ij}-\beta_{ij})}{(\alpha_{ij}-X_{ij})/(\alpha_{ij}-\beta_{ij})} \tag{1}$$

式(1)中反映了各指标达到目标的满意程度,X_{ij} 对证据链构成的贡献作用由 u_{ij} 表示,u_{ij} 的取值范围为 $[0,1]$,u_{ij} 趋于 1 为最满意,趋于 0 为最不满意。

构成证据链上的原子事件间存在先发事件的输出要素与后发事件的输入要素之间的因果关系。后发事件的输入要素与先发事件的输出要素构成了证据链的序数参量,设为 U_1、U_2,一般采用集成方法来计算各个序参量有序程度的总贡献度。具体方法为:

$$U_A(U_i)=\sum^{n}\lambda_j u_{ij},\ \sum_{j=1}^{n}\lambda_j=1,\lambda_j>0 \tag{2}$$

其中,$U_A(U_i)$ 为某个原子事件对事件证据链系统的总序参量;λ_j 为影响因素指标的权重。

(2)关联度函数。

原子事件相互作用的关联度模型可以表示为:

$$C=\left[\frac{U_A(U_1)U_A(U_2)\cdots U_A(U_i)}{(U_A(U_1)+U_A(U_2)+\cdots+U_A(U_i))(U_A(U_1)+U_A(U_2)+\cdots+U_A(U_i))}\right]^{1/2} \tag{3}$$

其中,C 表示事件之间的关联度;i 表示事件个数;$U_A(U_i)$ 表示对证据链系统有序的作用贡献大小。

当 C 趋于 1,事件关联程度最大,即若干原子事件的演化将会对证据链产生完全的影响;当 $C=0$ 时,表示事件之间无任何关联性。在计算过程中,设定一个关联度的临界值 α,若 $C>\alpha$,表示后发事件的输入可以由先发事件的输出表示,即后发事件的发生由先发事件引起;否则事件间不存在链式作用关系。

设集合 $E=\{e_1,e_2\cdots e_i\cdots e_n\}$,通过计算事件关联度,可以构建一条以初始原子事件为链源的证据链,具体过程为:

设初始原子事件为 $e_i(e_i\in E)$,以 e_i 为先发事件,E 中的其他事件为后发事件,计算 e_i 与其他事件的关联度 $C_{ij}(1\leqslant j\leqslant n)$。当 $C_{ij}>\alpha$ 时,则说明 e_i 与 e_j 之间的关联度较高,它们之间存在着潜在的链式关系。令 e_i 存在潜在的链式关系的后继事件集合由 $E_{il}=$

$\{e_{i1},e_{i2}\cdots e_{ii}\cdots e_{in}\}$ 表示。接着按照同样的方法,对集合 $E_{il1},E_{il2},\cdots,E_{ili},\cdots,E_{iln}$ 进行操作。其中 E_{ili} 表示与集合 E_{il} 中的事件 e_{ii} 具有链式关系的事件集合。以上过程持续直到在集合 E 中不再找到与后继事件集合中的事件具有链式关系的原子事件为止。最后,从初始事件 e_i 开始,合并所有的后继原子事件集合 $E_{il},E_{ili},E_{ilii},\cdots$,可以得到最后的目标证据链 $EL=\{E,P\}$。其中 $E=\{e_1,e_2,\cdots,e_i,\cdots,e_m \mid 1\leqslant m\leqslant n\}$ 为原子事件集合;$P=\{(e_i,e_j)\mid e_i,e_j\in E\}$ 为 E 中各种原子事件的链式关系集合。

2.3　证据链的形式化表示

设 $E=\{e_i\mid 1\leqslant i\leqslant m\}$ 为事件的输入变量集合,即证据事件序列,$A=\{a_k\mid 1\leqslant k\leqslant m\}$ 为控制变量集合,$S=\{s_j\mid 1\leqslant j\leqslant n\}$ 为状态变量集合,$Z=\{z_j\mid 1\leqslant j\leqslant m\}$ 为事件承载状态变量集合,$O=\{o_j\mid 1\leqslant j\leqslant m\}$ 为输出变量集合,s^e 为事件的触发状态变量。

用函数 begin(p)表示初始路径,用 end(p)表示路径尾,基本的状态转换路径表示为 P_φ。设有两条路径:

$$p_x=(s_{x_1},e_{x_1},s_{x_2},e_{x_2},\cdots,s_{x_m},e_{x_m})$$

$$p_y=(s_{y_1},e_{y_1},s_{y_2},e_{y_2},\cdots,s_{y_n},e_{y_n})$$

如果 $\mathrm{end}(p_x)=\mathrm{begin}(p_y)$,如 $s_{y_m}=s_{y_1}$,则两条路径可以连接为一条路径 p。即:

$$p=p_x\otimes p_y=(s_{x_1},e_{x_1},s_{x_2},e_{x_2},\cdots,s_{x_m},e_{x_m},s_{y_1},e_{y_1},s_{y_2},e_{y_2},\cdots,s_{y_n},e_{y_n}) \tag{3}$$

其中,符号“$\otimes$”表示连接操作。

将 $(s_{x_1},s_{x_2},\cdots,s_{x_m},s_{y_1},s_{y_2},\cdots,s_{y_n})$ 表示为 p 的状态序列,$(e_{x_1},e_{x_2},\cdots,e_{x_{m-1}},e_{y_1},e_{y_2},\cdots,e_{y_n})$ 表示 p 的事件序列。M 表示相应路径 p 的最大消耗时间,m 为最大消耗时间,p 的持续时间定义为 $p^{\mathrm{T}}=[m,M]$。设定一个时间参照点 v,路径 p 的时间伴随属性可转换为两个绝对的时间区间 $[v,v+m]$ 和 $[v,v+M]$。P^E 为输入符号串,P^O 为输出符号串,设 $p=(s_i,e_y,s_j)$,$p_x^o=\mu(s_i,e^y)=\delta_k$,$p_x^{\mathrm{T}}=\tau(s_i,e_y)=[m_x,M_x]$,$\mu:S\times E\rightarrow$ 输出函数,$\tau:S\times E\rightarrow R_0\times(R_0\cup\{\infty\})$。并且 $p_x^e=(e_y)$,$e_y\in E$,单状态下 $p^{\mathrm{T}}=[0,0]$。因此式(4)更新为:

$$p=\sigma(p_x,p_y)=\begin{cases}p=p_x\otimes p_y\\ p^e=p_x^e\otimes p_y^e\\ p^o=p_x^o\otimes p_y^o\\ p^{\mathrm{T}}=p_x^{\mathrm{T}}+p_y^{\mathrm{T}}=[m_x+m_y,M_x+M_y]\end{cases} \tag{5}$$

对应路径集合连接运算为:

$$P=\varphi(P_1,P_2)=\{\sigma(p_1,p_2)\mid p_1\in P_1,p_2\in P_2,\mathrm{end}(p_1)=\mathrm{begin}(p_2)\} \tag{6}$$

3　拟合缺失数据

在取证研究中,获取的证据源数据并不都是完整的,存在犯罪人员将数据擦除或者篡改的危险,造成数据缺失,为此需要将缺失数据进行完整拟合。贝叶斯网络[12]可以自然地表示因果信息,是一种表示变量集合的连接概率分布的图形模型,在处理带有噪声和不完整数据集方面具有优势。该模型采用概率测度的权重来描述数据间的相关性。文中将缺失数据作为网络节点,数据间的因果关系采用有向图表示,进而构建贝叶斯网络结构。

3.1　贝叶斯网络

贝叶斯网络描述由两部分组成:

(1)有向无环图(DAG),其中每一个节点代表一个数据变量 X_i,P_{a_i} 为 X_i 的父节点的集合。

(2)条件概率表(CPT),表中的每一元素为数据变量 X_i,条件概率密度为 $p(X_iP_{a_i},\theta)$。

这两部分确定了贝叶斯网络,节点变量可以是对任何问题的抽象,文中节点变量主要指与原子事件发生、发展相关的各种因素。

3.2　证据链缺失数据的分析处理

基于贝叶斯统计的缺失证据参数学习[13]的基本思想是:对于一个随机变量 λ,服从先验分布 $P(\lambda)$,该分布表示学习前关于参数 λ 的先验信息。假设 $P(\lambda)$ 是一个均匀分布。参数 λ 的信息随着在实例数据集合 M 上的学习而发生变化。一般参数的估计值采用最大后验分布。采用式(7)计算参数的估计值为:

$$\lambda_{ijk}=\frac{\alpha_{jk}+N_{ijk}}{\sum_{k=1}^{r}(\alpha_k+N_{ijk})} \tag{7}$$

其中,α_k 代表先验知识(专家证据集[9]),特殊情况下,假设变量取各个值的概率都相等,即 $\alpha_i=1$,一般采用等价抽样规模法进行估计。

原子事件的贝叶斯网络拟合:令 $G=\{N,E,P\}$ 为原子事件贝叶斯网络,如图 1 所示。

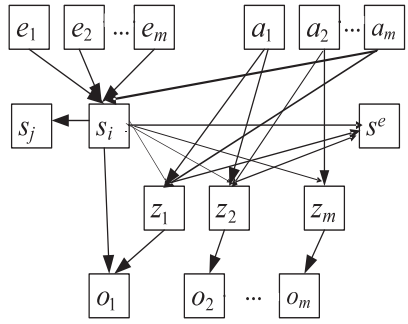

图 1　原子事件贝叶斯网络结构

其中,$N=Z\cup S\cup O$,(N,E) 表示网络结构,其作用是描述变量之间的因果关系,用条件概率表示变量

之间的影响程度,根据历史数据或通过专家知识直接指定得到变量的条件概率。得到其他节点的条件概率和根节点的先验概率,就可以得到所有变量的联合概率分布,如式(8):

$$p(e_i,s_i,s_j,z_j,s^e,a_k,o_j)=p(e_i)p(s_j\mid e_i)p(s_j,z_j\mid a_k)p(s_j\mid s_i)p(z_j\mid s_i)p(s^e\mid s_i,z_j)p(o_j\mid s_i)p(o_j\mid z_j) \tag{8}$$

通过式(8) 可以得到网络中各节点的边缘概率,确定先验网络。假设获取的部分信息为 E,利用此数据更新网络中其他节点的概率,实现对证据事件后果的预测和关键状态,由贝叶斯公式计算如下:

令 $e \in E$ 为证据信息,则:

$$p(s_i,z_j,s^e,o_j\mid e)=\frac{p(s_i,z_j,s^e,o_j,e)}{p(e)} \tag{9}$$

当网络节点过多时,为了降低计算复杂度,可以采用联合树推理算法[14]进行求解。

对缺失证据事件的修补,为取证提供完整证据链,以满足电子证据的分析需求。而缺失的原子事件又是离散的,因此,可以构造一种用于多个离散变量的贝叶斯网络。

4 测试结果与分析

以一次关联实验为例,测试该方法构造证据链关联事件的性能。测试环境为实验室内局域网(25 台主机,1 台服务器),操作系统为 Windows XP。数据采集了 2 500 个事件,测试中时间阈值(即在某一时间段内进行关联分析)分别为 20 min,40 min,60 min。实验中关联度临界值 α 设为 0.7。测试结果如图 2 所示。

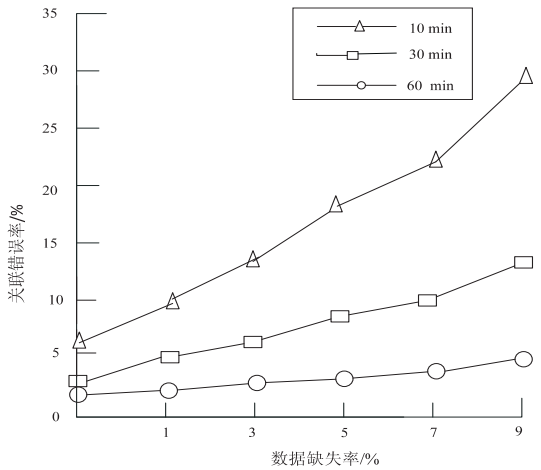

图 2 证据链关联结果

从测试结果中得出,时间阈值较小时,由于获取的前后知识不充分,错误率较大,随着阈值的增大,错误率明显下降。当数据缺失较严重时,错误率增加不明显,说明该方法对于缺失数据的拟合补充效果较为明显。但是当时间阈值较小,如 20 min 时,错误率却较高,说明此时对于因果关系的分析还不充分,仍需要进一步的自学习。

经过证据链中的事件关联分析后,减少了无用知识与冗余事件,证据分析的数据量减少了许多,见表 1,使得取证分析更有针对性。

表 1 事件数量比较

事件个数	关联后事件个数	精简比	关联前事件种类	关联后事件种类
2 500	893	2.8:1	37	21

5 结束语

文中引入关联度的概念描述证据事件间的相互影响程度,提出一种基于事件关联的证据链构造方法,综合不同的证据事件源进行相关性分析,去除冗余事件,最终构成证据链,有效地实现了电子证据的形式化表示,减少了证据分析的数据量。运用贝叶斯网络推理算法分析缺失数据与现有数据之间的因果关系,即使在部分事件失序和数据缺失情况下,算法也可推理犯罪入侵的发生过程,拟合证据链,保证了数据的完整性。形成证据链后,不仅能有效验证证据的原始性,而且能防止对证据记录的破坏,最大程度地保护证据,满足了电子取证的事件连续性的原则。但是随着网络取证的数据量的增大,特别是云计算技术的发展,给电子取证技术带来了挑战,比如构造海量数据的证据链,海量信息的证据事件处理,以及多维证据的分析等,这将是下一步研究的方向。

参考文献:

[1] Han J,Kamber M,Pei J. Data mining concepts and techniques [M]. 3nd ed. Beijing:China Machine Press,2012:288-293.

[2] Ding L P,Wang Y J. Study on relevant law and technology issues about computer forensics[J]. Journal of Software,2005,16(2):260-275.

[3] Etzion O,Niblett P. Event processing in action[M]. [s. l.]: Manning Publications Co. ,2010.

[4] Ning Peng,Cui Yun,Reeves D S. Analyzing intensive intrusion alerts via correlation[C]//RAID 2002. Zurich,Switzerland: [s. n.],2002.

[5] Koch G G,Koldehofe B,Rothermel K. Cordies:expressive event correlation in distributed systems[C]//Proceedings of the fourth ACM international conference on distributed event-based systems. [s. l.]:ACM,2010:26-37.

[6] Acampora G. Exploiting timed automata based fuzzy controllers for designing adaptive intrusion detection systems[J]. Soft Computing,2012,16(7):1183-1196.

[7] Tiffany M. A survey of event correlation techniques and related

(下转第 116 页)

运算的次数比起 Shamir 算法要少很多。Shamir-ZLMOF 算法的点加运算次数与 Shamir-NAF 算法相同，但是倍点运算次数大多数情况下要少。因此 Shamir-ZLMOF 算法比起以上两种算法具有更高的执行效率。

5 结束语

文中首先分析了当前主流的标量乘和多标量乘算法。在对 MOF 算法研究的基础上利用左移性质提出了 ZLMOF 算法，然后利用 Homer 规则采用连续倍点运算，进一步提高算法的效率。ZLMOF 算法的倍点运算效率比起像 NAF 等传统算法明显要高。紧接着结合滑动窗口法提出了 ZLMOF—滑动窗口算法，采用预计算进一步提高运算速率。同时，结合 Shamir 算法和 ZLMOF 算法提出了 Shamir—ZLMOF 算法。新算法比传统的 Shamir—NAF 算法效率要高。

参考文献：

[1] Diffie W, Hellman M. New directions in cryptography[J]. IEEE Transactions on Information Theory, 1976, 22(6): 644-654.

[2] Rivest R L, Shamir A, Adleman L M. A method for obtaining digital signatures and public key cryptosystem[J]. Communications of the ACM, 1987, 21(2): 120-126.

[3] Koblitz N. Elliptic curve cryptosystems[J]. Mathematics of Computation, 1987, 48(177): 203-209.

[4] Miller V S. Use of elliptic curves in cryptography[C]//Advance in cryptology. [s. l.]: Springer, 1986: 417-426.

[5] Huang X, Shah P, Sharma D. Fast algorithm in ECC for wireless sensor network[C]//Proceedings of the international multi conference of engineers and computer scientists. [s. l.]: [s. n.], 2010: 17-19.

[6] Balasubramaniam P, Karthikeyan E. Elliptic curve scalar multiplication algorithm using complementary recoding[J]. Applied Mathematics and Computation, 2007, 190(1): 51-56.

[7] Okeya K. Signed binary representations revisited[C]//Proceedings of CRYPTO'04. [s. l.]: [s. n.], 2004: 123-139.

[8] Kodali P K, Budwal H S. High performance scalar multiplication for ECC[C]//International conference on computer communication and informatics. Piscataway: IEEE, 2013: 1-4.

[9] Solinas J A. An improved algorithm for arithmetic on a family of elliptic curves[C]//Advances in cryptology. [s. l.]: [s. n.], 1997: 357-371.

[10] Shah P G, Huang X, Sharma D. Sliding window method with flexible window size for scalar multiplication on wireless sensor network nodes[C]//International conference on wireless communication and sensor computing. [s. l.]: IEEE, 2010: 1-6.

[11] Solinas J A. Low-weight binary representations for pairs of integers[R]. [s. l.]: Centre for Applied Cryptographic Research, 2001.

[12] Zhang J Z, Kou Y Z, Wang T, et al. Fault analysis on elliptic curve cryptosystems with sliding window method[J]. Journal on Communications, 2012, 33(1): 71-78.

[13] Yong K L, Ingrid V. A compact architecture for Montgomery elliptic curve scalar multiplication processor[M]. [s. l.]: Springer, 2007: 115-127.

(上接第 110 页)

topics[EB/OL]. 2003. http://www.tiffman.net/netman/netman.pdf.

[8] Jakobson G, Weissman M. Real-time telecommunication network management: extending event correlation with temporal constraints[C]//Proceedings of the fourth symposium on integrated network management. Santa Barbara, California, USA: Chapman & Hall, 1995: 290-301.

[9] Narayanan K, Bose S K, Rao S. Towards' integrated monitoring and management of DataCenters using complex event processing techniques[C]//Proceedings of the fourth annual ACM conference. Bangalore: ACM, 2011.

[10] Luckham D. The power of events: an introduction to complex event processing in distributed enterprise systems[M]. [s. l.]: Addison-Wesley, 2002.

[11] 张有东，曾庆凯，王建东. 网络协同取证计算研究[J]. 计算机学报，2010，33(3)：504-513.

[12] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference[M]. San Mateo, CA: Morgan Kaufman Publishers, 1988.

[13] Cheng J, Russell G, Kelly J. Learning Bayesian networks from data: an information-theory based approach[J]. Artificial Intelligence, 2002, 137(1-2): 43-90.

[14] Gui H X. Research of the intrusion detection system based on data mining[C]//Proceedings of the international conference on e-education entertainment and e-management. [s. l.]: [s. n.], 2011: 190-192.