(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109918505 B
(45)授权公告日 2023.09.19

(21)申请号 201910141724.0

(22)申请日 2019.02.26

(65)同一申请的已公布的文献号
申请公布号 CN 109918505 A

(43)申请公布日 2019.06.21

(73)专利权人 西安电子科技大学
地址 710071 陕西省西安市太白南路2号
专利权人 中国科学院大学

(72)发明人 姬逸潇 张玉清

(74)专利代理机构 北京君尚知识产权代理有限公司 11200
专利代理师 司立彬

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/951(2019.01)
G06F 16/958(2019.01)

(56)对比文件
CN 107733693 A,2018.02.23
CN 104091038 A,2014.10.08
姬逸潇等.基于超网络的网络安全事件连锁演化模型.《信息安全学报》.2019,
Zhenyan Liu等.Machine Learning for Analyzing Malware.《Journal of Cyber Security》.2017,
李再华等.基于特征挖掘的电网故障诊断方法.《中国电机工程学报》.2010,(第10期),

审查员 何华

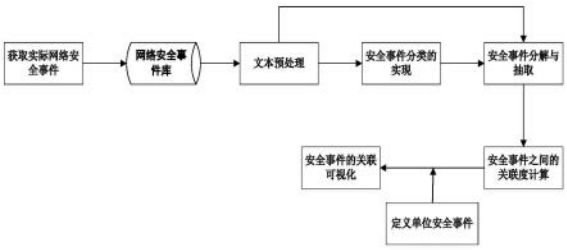
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种基于文本处理的网络安全事件可视化方法

(57)摘要

本发明公开了一种基于文本处理的网络安全事件可视化方法,其步骤包括:1)获取多个网络安全事件生成一事件集合,并将其转化为一结构化特征向量;2)对每一网络安全事件进行分类标注;3)将每一网络安全事件中的分词按设定的抽取目标进行注释,然后根据注释抽取各网络安全事件的安全事件内容;4)根据每一网络安全事件的安全事件内容,构建出一基础FP树,从该基础FP树中获得条件模式基,利用该条件模式基,再构建一个新的条件FP树,直至新的条件FP树中仅包含一个元素项,得到该网络安全事件的关联频繁集;5)根据关联频繁集,确定对应网络安全事件的关联度强弱以及各网络安全事件之间的连接关系,对该事件集合进行可视化。



CN 109918505 B

1.一种基于文本处理的网络安全事件可视化方法,其步骤包括:

1)获取多个网络安全事件生成一事件集合,并将该事件集合转化为一结构化特征向量;生成所述结构化特征向量的方法为:首先对每一所述网络安全事件进行分词,然后计算每一分词的TF/IDF权值,将分词TF/IDF权值大于设定阈值的分词作为特征词;然后对选出的各特征词设置编号并依据编号将各个特征词的TF/IDF权值按行写入矩阵,得到所述结构化特征向量;其中,每一网络安全事件对应一段描述该网络安全事件的文字信息;

2)利用分类模型对每一所述网络安全事件进行分类标注,分类标注方法为:首先使用梯度下降算法在代价函数的参数空间中搜索所述结构化特征向量中不同特征TF/IDF权值的最优解,然后根据特征的TF/IDF权值及其最优解利用反向传播算法来计算权值导数,用于计算特征与类别间联合概率分布;然后对于每一待标注的网络安全事件,依据各特征与类别间联合概率分布,分别计算在输入该待标注网络安全事件的特征时,该待标注网络安全事件属于每个类别的后验概率,将具有最大后验概率的类别k作为该待标注网络安全事件的分类预测结果;

3)将每一网络安全事件中的分词按设定的抽取目标进行注释,然后根据注释抽取各网络安全事件的安全事件内容;

4)根据每一网络安全事件的安全事件内容,构建出该网络安全事件的基础FP树,从该基础FP树中获得条件模式基,利用该条件模式基,再构建一个新的条件FP树,然后判断当前新的条件FP树中是否仅包含一个元素项,如果不是,则从当前新的条件FP树中获得条件模式基,利用条件模式基,再构建一个新的条件FP树,直到新的条件FP树包含一个元素项为止;得到该网络安全事件的关联频繁集;

5)根据网络安全事件的关联频繁集,确定对应网络安全事件的关联度强弱以及各网络安全事件之间的连接关系,对该事件集合进行可视化。

2.如权利要求1所述的方法,其特征在于,步骤5)中,对该事件集合进行可视化的方法为:

51)将该事件集合中的每一安全事件抽象为一节点,初始化超网络只存在一节点a以及与该节点a连接的n条超边;其中,存在直接因果关系的节点间通过超边连接,每一节点包含若干单位安全事件,所述单位安全事件为安全事件的原因或安全事件的结果;

52)每次选取一未加入超网络的节点i并将其加入到该超网络,并根据节点i与超网络当前存在的节点之间的关系建立节点间的超边连接,根据该事件集合对应的关联关系;

53)根据步骤52)的处理结果得到该事件集合对应的关联关系,并在超边中加入箭头表示因果关系,对该事件集合进行可视化。

3.如权利要求2所述的方法,其特征在于,步骤52)中,从超网络已有的节点中的按照概率优先选取待加入超网络的节点。

4.如权利要求3所述的方法,其特征在于,步骤52)中,根据公式$\prod d_H(i)=\frac{d_H(i)}{\sum_j d_H(j)}$计算各未加入超网络的节点i的概率$\prod d_H(i)$;其中,$d_H(i)$为节点i的超度,$d_H(j)$为超网络中节点j的超度。

5.如权利要求1所述的方法,其特征在于,设定的抽取目标包括:漏洞类型、受影响产品

以及漏洞利用方式;依据抽取目标设定若干注释,每一注释对应一数字编号。

6.如权利要求1或5所述的方法,其特征在于,依据抽取目标设定的注释包括:产品修饰词、受影响产品、漏洞、漏洞利用方式、背景词、触发词、攻击者、助词或情态词、漏洞修饰词和利用方式修饰词。

7.如权利要求6所述的方法,其特征在于,步骤3)中,将网络安全事件分词得到的词汇序列为隐马尔可夫模型中的观测序列,将各个词汇序列对应的标注序列为隐马尔可夫模型的状态序列;然后采用基于触发关键字的规则抽取该网络安全事件中的关键词,得到该网络安全事件的安全事件内容。

8.如权利要求1所述的方法,其特征在于,根据网络安全事件的关联频繁集中关键字的数量,确定对应网络安全事件的关联度强弱。

一种基于文本处理的网络安全事件可视化方法

技术领域

[0001] 本发明属于网络安全技术领域，更进一步涉及一种基于文本处理的网络安全事件可视化方法。本发明主要用来实现安全事件语义关联关系模型的建立，揭示安全事件连锁反应机理，为提出安全事件描述框架及度量指标体系奠定基础。所述模型可适用于不断更新的网络安全事件的及时处理与响应，也对攻防技术发展趋势分析与预测有极其积极的意义。

背景技术

[0002] 对网络安全(Network Security)的评估与度量指的是提取一定时间、一定空间内的网络安全相关的要素进行分析，针对安全数据进行综合处理，确定系统受到的攻击行为，提供网络安全的整体情况，进而评估网络系统安全状态，并基于分析结果预测其未来的发展趋势。

[0003] 随着计算机技术和通信技术的迅速发展以及用户需求的不断增加，计算机网络规模日益庞大，应用系统日益复杂。网络安全威胁的范围和内容不断扩大和演化，网络安全形势与挑战日益严峻复杂。如何全方位感知网络安全态势、实时监控网络运行状况、保障信息资产安全，应该引起我们足够的重视。因此，针对网络安全评估模型及关键技术已经成为目前网络安全领域的研究热点。

[0004] 由于网络安全事件层出不穷，针对于安全事件的防御技术和网络攻击危害的预测具有十分重要的意义。网络安全事件可以反映出的技术和危害之间存在的内在关联。网络安全事件带来的危害，对人们的各方面的生活造成不同程度的影响。道高一尺，魔高一丈，安全事件带来危害的同时，也激励着技术的进一步发展，不断出现的新技术或新方法来应对各种网络攻击带来的危害。

[0005] 通过对目前安全事件研究现状的分析，可以得出现在的安全事件分析方法存在以下三点不足：

[0006] (1)对实际网络安全事件的处理没有提出一套完整的处理方案，仅仅针对于不同安全事件的不同方面去提出应对措施。

[0007] (2)对网络安全事件的研究主要集中在对响应模型的提出，并没有对安全事件之间的内在关联的机制和原理进行进一步揭示，即安全事件时空语义的关联关系。

[0008] (3)主要的安全事件数据来源集中在入侵检测数据和日志数据等方面，并没有针对于网络安全事件的文字数据的分析和研究。

发明内容

[0009] 本发明的目的在于提供一种基于文本处理的网络安全事件可视化方法，能够将实际网络安全历史事件转化为网络安全度量与评估的重要参数。

[0010] 本发明基于文本处理的网络安全事件可视化方法，包括如下步骤：

[0011] 1)获取实际网络安全事件：通过调研国内外知名安全资讯网站，确定具有研究意

义的某些网站，通过编写网络爬虫进行网络安全事件文字信息的收集；

[0012] 2)文本信息预处理：首先是进行数据清洗，即去除文本中的特殊字符、统一文本格式，之后将清洗后的文本转化为由数字表示的结构化特征向量，即生成网络安全事件样本的样本向量并将样本向量矩阵化，从而使得算法可以对其进行解析，结构化特征向量的生成主要可分为中文分词、特征提取以及向量化表示三个部分；

[0013] 3)网络安全事件分类实现：

[0014] (3a)文本类别确定：依据中国国家标准化管理委员会发布的《信息安全事件分类分级指南》，综合本方法的实现目标与现今的网络安全形势，决定将事件依据其威胁形式划分为五个大类，分别是：漏洞预警事件、恶意软件事件、信息泄漏事件、网络攻击事件及其他安全事件；

[0015] (3b)神经网络分类模型对每一样本(即上述采集的安全事件)进行分类标注：对于未知的事件样本，模型的输入参数为该样本中每个特征词的TF/IDF权值，首先使用梯度下降算法在代价函数的参数空间中搜索结构化特征向量中不同特征TF/IDF权值的最优解，然后根据特征的TF/IDF权值及其最优解利用反向传播算法来计算权值导数，用于计算特征与类别间联合概率分布；最后依据训练时得到的各特征与类别间联合概率分布(通过回归算法使神经网络的输出权值转化为概率，得到结构化特征向量与类别间的联合概率分布)，分别计算在输入该未知样本的特征时(即该未知样本的特征词TFIDF权值时)，该未知样本属于五个类别的后验概率，其中具有最大后验概率的类别k即为未知样本的分类预测结果。

[0016] (3c)Adaboost算法对分类模型的优化：首先初始化权值矩阵，训练得到弱分类器并进行测试，之后将测试结果中被正确分类的样本的权重降低，将测试结果中被错误分类的样本的权重提高，将每一轮得到的弱分类器依据其准确率计算分类器系数，最终使用该系数将各个弱分类器组合为一个强分类器，完成分类模型的优化。

[0017] 4)安全事件分解与抽取：以步骤2)中的文本预处理工作为基础，先将分词后的训练样本按抽取的目标进行注释，然后将训练样本中的中文句子转换为数字向量形式以方便模型算法进行分析处理，最后采用基于触发关键字(关键词为注释中的“标记状态”)的规则抽取与隐马尔可夫模型相结合的思路，进行安全事件内容的抽取；

[0018] 5)安全事件之间的关联度计算：根据步骤4)中从每一安全事件抽取出的关键字，将关键词作为FP-growth算法的输入参数，构建出该安全事件的基础FP树，从该基础FP树中获得条件模式基，利用该条件模式基，再构建一个新的条件FP树，迭代重复步骤1步骤2(从当前新的条件FP树中获得条件模式基，利用条件模式基，再构建一个新的条件FP树)，直到当前新的条件FP树包含一个元素项为止，即可得到关联频繁集。然后根据不同安全事件通过上述方法得到的关联频繁集中关键字的数量，来判断关联度的强弱，即某个安全事件的关联频繁集中关键字数量越多，则该安全事件的关联度越强。

[0019] 6)定义单位安全事件：单位安全事件(也称为原子安全事件或简单安全事件)是指在网络空间环境中，以微观角度直接观察到的、最基本的不能再分解的安全事件，任何安全事件从宏观角度都可以表示为若干个单位安全事件的并集集合。一个单位安全事件可以是某一个安全事件的原因，也可以是某一个安全事件的结果。

[0020] 7)安全事件的关联可视化：根据步骤5)中得到的关联度计算结果，进行不同安全事件之间强弱关系的定义，定义强弱关系区分的阈值。结合BA无标度网络演化模型的算法

以及系统论中超网络的概念，提出一种基于超网络的网络安全事件连锁演化模型，模型建立步骤如下：

[0021] (7a)初始化：将具体安全事件抽象为节点a,b,c,d,e,f…有限个节点。默认开始超网络只存在节点a，以及包含着与这个节点a连接的n条超边，n为自然数，超边连接与a存在直接因果关系的安全事件。

[0022] (7b)超边增长：每次增加一个新的节点，根据新加入节点与超网络当前存在的节点之间的关系建立节点间的超边连接。

[0023] (7c)优先连接：从a节点开始，不断加入其他节点，并从已有的超网络中的节点按照概率优先选取节点，与新加入的节点结合生成超边。根据公式$\prod d_H(i)=\frac{d_H(i)}{\sum_j d_H(j)}$，每次选取连接的节点i的概率$\prod d_H(i)$等于节点i的超度$d_H(i)$(节点的超度定义为包含该节点的超边个数)与超网络中的已有节点j的超度$d_H(j)$总和之比，计算每次选中某个节点i的概率，i可以是a,b,c,d,e,f…中任意一个节点。最后可以得到每个节点的超边数量。

[0024] (7d)根据最终的节点和超边的数量，得到事件层的关联关系，并根据实际情况在超边中加入箭头表示因果关系，同时在基础设施层(根据安全事件以及超网络多层次性的特点，可将基于安全事件的超网络模型定义为两个层次，即事件层和基础设施层，基础设施层涉及实际基础设施、地域环境以及传播载体等不同的参数因素，可以使关联可视化的结果更为)加入实际基础设施、地域环境以及传播载体等不同的参数因素。

[0025] 在本发明一个较佳实施例中，所述步骤1)中，获取不同种类的网络安全事件文本信息有45000～50000条。

[0026] 在本发明一个较佳实施例中，所述步骤4)中，预处理过后训练样本共有9952个特征维度，事件共计被划分为5个类别，构造神经网络的输入层神经元数目为9952，隐藏层深度为1，隐藏层神经元数目为10，输出层神经元数目为5，模型的初始学习速率为1.5。

[0027] 本发明的有益效果是：

[0028] 1)采用本发明的可视化方法，为网络安全的评估与度量提供了新的参数基础，运用自然语言处理的相关方法对网络安全性进行了全新的描述，解决了之前安全事件研究大多只面向日志数据和IDS数据的缺陷；

[0029] 2)采用深度学习Adaboost算法，针对本发明需要处理的安全事件文本信息，与普遍使用的朴素贝叶斯分类模型相比，更加重视文本的语义特征，本发明提出的分类模型能够逐步发现出不同词汇之间的隐藏关系，从而使分类准确度大大提高；

[0030] 3)采用系统论的概念，结合超网络的特点，建立安全事件语义的关联关系模型，揭示安全事件连锁反应机理，为实现安全事件危害效用度量与评估方法提供理论基础，丰富网络安全评估领域的研究，使网络安全事件内在的联系可以清晰的呈现出来，从安全事件之间的关联关系入手，有助于建立安全事件对网络系统安全程度的影响分析，也有助于网络真实攻击和防御历史事件的分析以及攻防技术发展趋势分析；

[0031] 综上，本发明的可视化方法，具有分类精度高、模型创新性强、适用性强等突出特点。

附图说明

[0032] 图1是本发明一种基于文本处理的网络安全事件可视化方法的流程图。

具体实施方式

[0033] 下面结合附图对本发明的较佳实施例进行详细阐述，以使本发明的优点和特征能更易于被本领域技术人员理解，从而对本发明的保护范围做出更为清楚明确的界定。

[0034] 请参阅图1，本发明实施例包括：

[0035] 一种基于文本处理的网络安全事件可视化方法，包括如下步骤：

[0036] 1)获取实际网络安全事件：通过调研国内外知名安全资讯网站，确定可信度较高、专业性较强、信息更新及时的某些安全资讯网站。使用基于Python的Scrapy框架实现文本信息的获取，并借助selenium模拟浏览器环境完成了对由Ajax搭建的、页面异步加载的网站的信息抓取，并将所抓取的数据存入MySQL数据库，最终形成拥有43848条安全事件文本数据的网络安全事件库；

[0037] 2)文本预处理：首要是进行数据清洗，即去除文本中的特殊字符、统一文本格式，之后将清洗后的文本转化为由数字表示的结构化特征向量，从而使得算法可以对其进行解析，数据的结构化特征向量生成方法包括以下三个部分：

[0038] (2a)中文分词：采用基于词的划分方法，即将文本中的句子按其所包含的词汇进行分割，进而将整篇文档转换为由词语所组成的向量，用以描述其中蕴含的特征。

[0039] (2b)特征提取：使用了TF/IDF作为文本特征提取的方法，算法中的TF代表词频，表示特征词在整篇文本中出现的频率，IDF代表逆文本频率，表示特征词在所有训练样本中出现的频率。计算上述得到的每一分词(特征)的TF/IDF值，通过设立TF/IDF权值的阈值，过滤掉部分不具有特征意义的常见词，进一步降低模型的特征维度，避免发生过拟合。

[0040] (2c)向量化表示：为选出的特征词设置编号并依据编号将各个特征词的TF/IDF权值按行写入矩阵中作为训练样本矩阵。实施例模型训练共使用了2000个样本，得到了9952个特征值，构造的输入样本矩阵大小为2000×9952；

[0041] 3)网络安全事件分类实现：

[0042] (3a)文本类别确定：依据中国国家标准化管理委员会发布的《信息安全事件分类分级指南》，综合本方法的实现目标与现今的网络安全形势，决定将事件依据其威胁形式划分为五个大类，分别是：漏洞预警事件、恶意软件事件、信息泄漏事件、网络攻击事件及其他安全事件；

[0043] (3b)安全事件分类的实现：通过神经网络算法模型来实现安全事件的分类工作，模型的输入参数为该样本中每个特征词的TF/IDF权值，首先使用梯度下降算法来在代价函数的参数空间中搜索最优解，利用反向传播算法来计算权值导数，最后依据训练时得到的各个特征与类别间联合概率分布，分别计算在输入特征出现的前提下，样本属于五个类别的后验概率，其中具有最大后验概率的类别k即为未知样本的分类预测结果。实施例预处理过后训练样本共有9952个特征维度，事件共计被划分为5个类别，因此构造神经网络的输入层神经元数目为9952，隐藏层深度为1，隐藏层神经元数目为10，输出层神经元数目为5，模型的初始学习速率为1.5。

[0044] (3c)分类模型的提升与优化：考虑到安全事件的特点，将Adaboost算法应用于神

经网络分类模型。首先初始化权值矩阵，训练得到弱分类器并进行测试，之后将测试结果中被正确分类的样本的权重降低，将测试结果中被错误分类的样本的权重提高，将每一轮得到的弱分类器依据其准确率计算分类器系数，最终使用该系数将各个弱分类器组合为一个强分类器，完成分类模型的优化。

[0045] 4)安全事件分解与抽取：以步骤2)中的文本预处理工作为基础，需要以下三个具体步骤完成分解与抽取工作：

[0046] (4a)文本标注：文本标注的目的是将分词后的训练样本按抽取的目标进行注释，文本的标注序号需要尽可能地对句子中的各个成分予以区分。在实施例中，需要抽取的目标有：漏洞类型、受影响产品以及漏洞利用方式，依据抽取目标定义了10种标注状态

[0047]

状态号	含义	状态号	含义
0	产品修饰词	5	触发词
1	受影响产品	6	攻击者
2	漏洞	7	助词、情态词
3	漏洞利用方式	8	漏洞修饰词
4	背景词	9	利用方式修饰词

[0048] (4b)文本向量化：统计所有训练样本中的分词结果，将词汇去重后按序存入一个字典中，字典的键为中文词汇，值为该词汇被分配的序号，对于某些样本中的一些英文与数字序列，将这些随机序列用统一的特殊标识进行表示，最后使用字典将安全事件文本中的中文词汇转换为对应的序号，完成训练样本向量化。

[0049] (4c)基于隐形马尔科夫模型的关键字抽取：分词得到的词汇序列为模型中的观测序列，各个词汇序列对应的标注序列为模型的状态序列。在监督学习时，使用极大似然算法依据训练样本的人工标注结果计算模型的参数矩阵，完成模型的构建；

[0050] 5)安全事件之间的关联度计算：根据步骤4)中已经抽取出的安全事件的关键字，将关键词作为FP-growth算法的输入参数，构建出基础的FP树，从FP树中获得条件模式基，利用条件模式基，再构建一个新的条件FP树，迭代重复步骤1步骤2，直到树包含一个元素项为止，即可得到关联频繁集。然后根据不同安全事件通过上述方法得到的关联频繁集中关键字的数量，来判断关联度的强弱关系。

[0051] 6)定义单位安全事件：单位安全事件(也称为原子安全事件或简单安全事件)是指在网络空间环境中，以微观角度直接观察到的、最基本的不能再分解的安全事件，任何安全事件从宏观角度都可以表示为若干个单位安全事件的并集集合。一个单位安全事件可以是某一个单位安全事件的原因，也可以是某一个安全事件的结果。

[0052] 7)安全事件的关联可视化：根据步骤5)中得到的关联度计算结果，进行不同安全事件之间强弱关系的定义，定义强弱关系区分的阈值。结合BA无标度网络演化模型的算法以及系统论中超网络的概念，提出一种基于超网络的网络安全事件连锁演化模型，模型建立步骤如下：

[0053] (7a)初始化：将具体安全事件抽象为节点a,b,c,d,e,f…有限个节点。默认开始只存在节点a，以及包含着这个节点的n条超边，n为自然数，超边连接与a存在直接因果关系的安全事件。

[0054] (7b)超边增长：每次增加一个新的节点，与a节点结合生成新的超边。

[0055] (7c)优先连接:从a节点开始,不断加入其他节点,并从已有的超网络中的节点按照概率优先选取节点,与新加入的节点结合生成超边。根据公式$\prod d_H(i)=\frac{d_H(i)}{\sum_j d_H(j)}$,每次选取连接的节点i的概率$\prod d_H(i)$等于节点i的超度$d_H(i)$与超网络中的已有节点j的超度$d_H(j)$总和之比,计算每次选中某个节点i的概率,i可以是a,b,c,d,e,f…中任意一个节点。最后可以得到每个节点的超边数量。

[0056] (7d)根据最终的节点和超边的数量,得到事件层的关联关系,并根据实际情况在超边中加入箭头表示因果关系,同时在基础设施层加入实际基础设施、地域环境以及传播载体等不同的参数因素。

[0057] 以上所述仅为本发明的优选实施例而已,并不用于限制本发明,对于本领域的技术人员来说,本发明可以有各种更改和变化。凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

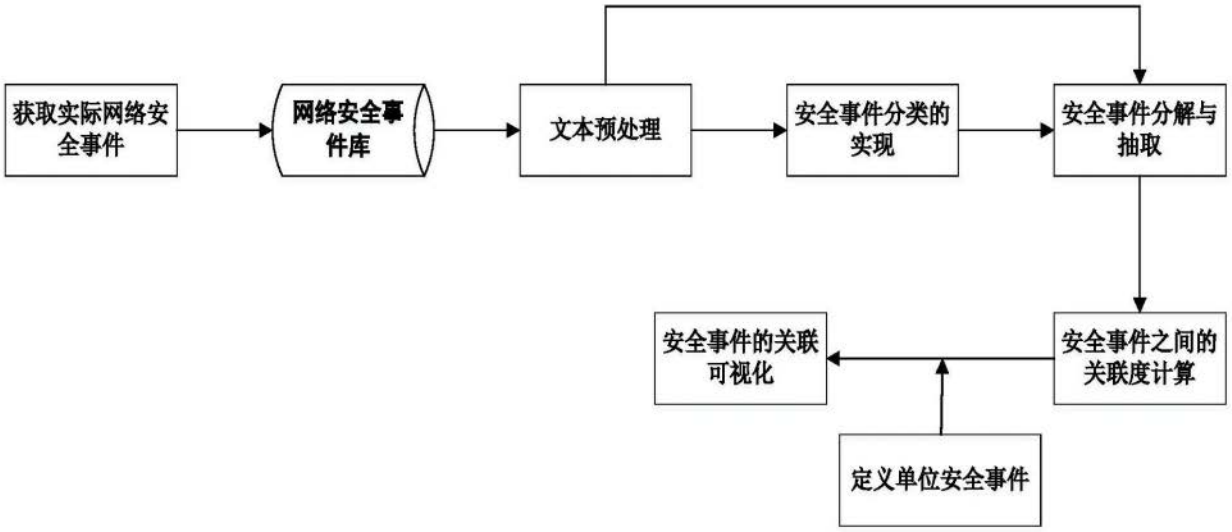

获取实际网络安全事件
网络安全事件库
文本预处理
安全事件分类的实现
安全事件分解与抽取
安全事件之间的关联度计算
安全事件的关联可视化
定义单位安全事件

图1