文章编号:1671-4598(2018)10-0304-05 DOI:10.16526/j.cnki.11—4762/tp.2018.10.065 中图分类号:TP393.08 文献标识码:A

大数据驱动下主动防御网络安全性评估技术

杨润佳
(哈尔滨理工大学,河北 秦皇岛 066003)

摘要:采用传统技术对安全性评估时受到外界干扰因素影响,使评估不精准,需引入网络熵对抗量化技术对大数据驱动下主动防御网络安全性进行评估,由此提高评估结果准确率;利用主动防御原理构建基于 Petri 网的安全性评估数学模型,模型中所体现攻击行为对网络造成间接影响,为降低间接影响造成的干扰需引入网络熵权衡收益,以提高网络使用安全性,通过权衡结果对主动防御网络安全性进行评估;实验结果表明,该评估技术准确率最高可达 96%,用户可在大数据驱动下使用网络,保障用户个人信息安全。

关键词:大数据驱动;主动防御;网络安全性;评估;网络熵;对抗量化

Security Assessment Technology of Active Defense Network Driven by Large Data

Yang Runjia
(Harbin Institute of Technology, Qinhuangdao 066003, China)

Abstract: When traditional technology is applied to the safety evaluation, it will be influenced by external interference factors, so that the evaluation is not accurate. It is necessary to introduce network entropy counter quantization technology to evaluate the safety of active defense network driven by big data, so as to improve the accuracy of evaluation results. Construction of Petri network security evaluation based on the mathematical model of the active defense principle, attack behavior reflects the model caused by indirect impact on the network, to reduce interference caused by indirect effects should be introduced to weigh the benefits in order to improve the network entropy, network security, by weighing the results of network active defense safety assessment. The experimental results show that the accuracy of the evaluation is up to 96%, and the user can use the network to ensure the security of the user's personal information under the large data drive.

Keywords: large data drive; active defense; network security; evaluation; network entropy; counter quantization

0 引言

主动防御网络仅仅采用防火墙虽然能够抑制攻击信息侵入,但是对于正常信息中夹杂的非法信息却无法识别,导致网络使用不安全、用户私密信息泄漏等现象,造成不必要的经济损失。为了保障用户在大数据驱动下使用网络同样安全,需对主动防御网络安全性进行评估。主动防御是一种具有深层次主动抵抗网络攻击行为的技术,可通过监测网络环境拦截外部非法入侵行为[1]。措施的采取也带来了一些安全性问题,对于主动防御的网络是否真正安全需要进行评估[2]。由于传统评估技术存在评估不精准问题,导致主动防御网络不安全,为此,提出了引入网络熵的模型对抗量化评估技术。通过与传统技术对比可知,该技术应用是具有合理性的。

1 主动防御网络安全性评估技术

1.1 主动防御原理

在大数据驱动下主动防御网络功能能够实现网络的安全运行,将检测技术与预测技术相结合,可保证网络使用的安全[3]。主动防御是在保证基本网络安全运行基础上实现的,除了传统系统防护技术外,还增加了响应技术,主动防御原理如图 1 所示。

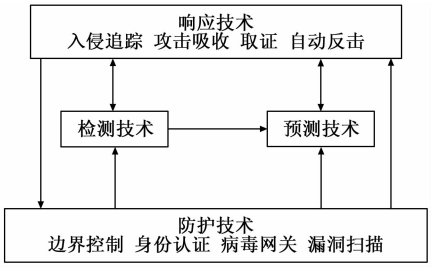

图 1 主动防御原理框图

依据传统防护技术,增加响应机制,并将检测原理与预测原理结合构成主动防御原理,以该原理为基础构建基于 Petri 网的安全性评估数学模型,在模型中引入网络熵权衡收益,降低间接影响造成的干扰,根据收益结果对主动防御网络安全性进行评估。

1.2 构建 Petri 网安全性评估数学建模

利用主动防御原理构建基于 Petri 网的安全性评估数学模型。采用 Petri 网建模方法可对整个主动防御网络安全性

收稿日期:2018-03-26; 修回日期:2018-04-21。

作者简介:杨润佳(1984-),男,河北承德人,工程师,主要从事电子信息工程方向的研究。

进行评估，从攻击者角度，将系统漏洞攻击主动防御网络行为进行模拟分析，并在必要安全信息基础上，查找潜在的多条主动防御安全性组合攻击途径[4]。

对 Petri 网主动防御安全性评估离不开对局部区域评估，在攻击场景下，依据评估结果实现评估模型的构建。要获取局部评估结果，需先收集与主动防御网络安全属性相关的数据，再结合这些数据分析网络的数据结构和存储结构，利用分析结果构建评估数学模型。

1.2.1　建模信息采集

扫描每台主机，搜寻主机中存在的已知安全性运行服务信息；对整个网络展开分析，获取各个主机间的联系，利用网络具有公开属性特点，通过扫描机可获取该特点的具体属性，以此为基础制定网络公开弱点的保护规则[5]。根据扫描获取的弱点信息以及制定的规则，提出 Petri 主动防御网络安全性评估方法；选择关联矩阵分析方式对数学模型构建展开讨论，分析出网络结构中存在的所有渗透序列，获取目标攻击路径，并计算成功可达概率。建模信息采集相关具体流程如图 2 所示。

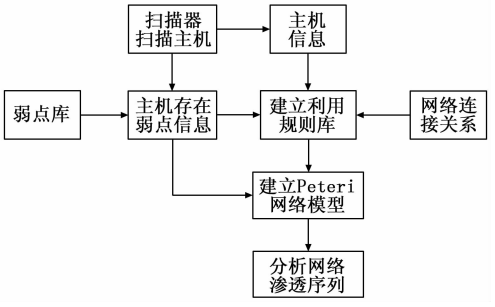

图 2　建模信息采集具体流程

由图 2 可知：结合以往攻击模型，对主动防御安全性展开说明，具体描述内容有：弱点描述、扫描机描述、网络连接关系描述和攻击规则描述。将网络弱点信息相关抽象数据组合成具体集合名称，方便安全性评估，具体集合名称如表 1 所示。

表 1　集合名称

名称	表示
host_name	安全缺乏所在主机名称
vulnerability_id	安全缺乏的唯一地址
vulnerability_ range = {local,remote}	安全缺乏所造成的攻击范围
vulnerability_service	安全缺乏利用主机的服务
vulnerability_result	安全缺乏所造成的攻击成功后在主机上获取相应权限
vulnerability_complexity	利用安全漏洞攻击的复杂程序

序利用安全漏洞攻击的复杂程序 complexity 进行归一量化处理：

1）不需要使用任何外界攻击工具，制定详细的攻击方案；

2）使用现有的攻击方案和工具；

3）缺乏攻击工具，那么需制定详细的攻击方案

4）公开主动防御网络漏洞，大致说明攻击方案；

5）公开主动防御网络漏洞，无攻击方案。

根据归一量化处理方案对用户等级进行划分：划分等级能够直观反映用户对于计算机的控制能力，对于某个主体来说，所有访问客体权限都是一个集合，通过分析 SANS 公布的排名前 10 网络漏洞弱点发现，涉及到的用户等级并不需要进行详细划分，通常为三个等级，分别是：Access、User、Root，其中 Access 代表远程接受网络服务，进行数据间的交互；User 代表经过管理员授予的用户权限，具有单独的空间与资源；Root 代表主机拥有所有资源控制权[6-8]。由于这三个等级之间关系都满足偏序分布 Access＜User＜Root，因此可一一对应到各个主机中。各个主机之间联系关系为：主机连接起点、源主机连接可达主机、源主机与普通主机之间无连系。

1.2.2　采集数据结构分析与存储

对采集到的建模信息，需进行详细分析。设立一串数据｛a，b，IP，TCP，HTTP｝可表示为：a 向 b 发送的 IP 数据是可达的，a 与 b 某个 TCP 端口是可以连接的，a 可对 b 进行 HTTP 服务。根据 TCP/IP 协议栈中各个层次的项目协议，需建立连接关系如表 2 所示。

表 2　协议栈层次连接关系

协议栈层次	连接关系
文本应用层	HTTP(超文本传输协议)、FTP(文件传输协议) SMTP(简单的消息传输协议)
协议传输层	TCP(终端控制协议)、UDP(用户数据报协议)
网络结构层	IP(网络互连协议)、ICMP(互联网控制报文协议)、IGMP(Protocol)
数据链路层	ARP(解决方案协议)、RARP(反向地址解析协议)

如果使用协议栈层次连接关系中的 TCP（终端控制协议）、HTTP（超文本传输协议）和 IP（网络互连协议）来约束链表 P _ T（place _ tansation，P _ T），那么在链表中只能存储来自数据库中的变迁集合，降低危险漏洞出现几率。使用链表 P _ T（place _ tansation，P _ T）来存储各个变迁集合，储存方式如图 3 所示。

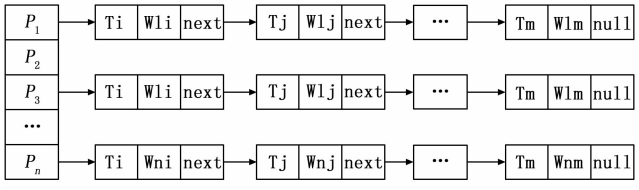

图 3　数据储存方式

在链表中 p 节点为 Petri 网安全性评估数学建模一个状态信息，该信息可由（ip，r）来表示，其中 ip 可代表主动防御受到到攻击位置，r 为主机所获取应用权限[9]。在链表

中 T 的节点为 Petri 网安全性评估数学建模的一个变迁行为信息，该信息可由（src，dst，vi）来表示，其中 src 可代表网络主动防御时受到攻击的标识信号，dst 为被攻击主机的标识信号，vi 为一次攻击行动的安全漏洞标识信号。P _ T 链表中的阈值范围代表了攻击级别，也代表了数据存储状态。

1.2.3 构建数学模型

设 $W=\{T_1,T_2\}$ 为一个集合，其中 T_1 为攻击者；T_2 为防御者；$S^k=(S_1^K,S_2^K,\cdots S_m^K)$ 为 T_k 的发生行为的空间，$S_i^1=\{(s_i^1,h_i^1,\pi_i^1)\mid 0<\pi_i^1<1,\sum\pi_i^1=1\}$ 为攻击者的攻击行为，s_i^1 为攻击者 T_1 的恶意行为 s_i；h_i^1 为攻击者 T_1 进行攻击的目标设备，该行为主要依赖主动防御技术和网络设备重视程度，根据设备种类和网络拓扑结构划分为 5 个攻击等级；π_i^1 为攻击者 T_1 进行恶意行为 s_i 的几率[10]。

$S_j^2=\{(s_j^2,h_j^2,\pi_j^2)\mid 0<\pi_j^2<1,\sum\pi_j^2=1\}$ 为防御者的防御行为，s_j^2 为防御者 T_2 的主动防御行为 s_j；h_j^2 为防御者 T_2 进行防御的目标设备，防御行为主要依赖网络结构，根据设备种类和网络拓扑结构划分为 5 个攻击等级；π_j^2 为防御者 T_2 进行主动防御行为 s_j 的概率。$Q=(Q_l\mid Q_l=(q_{l1},q_{l2},\cdots,q_{ln}))$ 为空间状态，n 为网络设备数量，Q_l 为在一定时间内的网络防御状态，$q_{li}=(host_{li},link_{lji})$，$host_{li}$ 为一定时间内网络设备获取权限状态，$link_{lji}$ 为网络设备发起入侵链路性能下降程度。不同状态代表不同网络防御阶段，状态之间相互转换是由防御等级与攻击等级共同决定的，依据防御与攻击双方行为可构建数学模型，如公式（1）所示：

$$q_{ij}^l=W_{ij}^l+\sum_{l=1}^{k}P_{ij}^l(Q_l\mid q_l,S_i^1,S_i^2)S_l,P_{ij}^l\geqslant 0,\sum_{l=1}^{k}P_{ij}^l<1 \tag{1}$$

公式（1）中，W_{ij}^l 为防御与攻击双方在 Q_l 状态下的行为；(S_i^1,S_i^2) 为防御者的收益；q_l 为攻击成功后的网络状态；$\sum_{l=1}^{k}P_{ij}^l(Q_l\mid q_l,S_i^1,S_i^2)S_l$ 为攻击行为对网络所产生的间接影响；$(Q_l\mid q_l,S_i^1,S_i^2)$ 为信息状态迁移几率。

1.3 引入网络熵权衡收益

由于在构建模型过程中受到外界攻击，导致网络产生间接影响，造成后续安全性评估出现干扰因素，严重扰乱评估结果，为此引入网络熵对抗量化技术来权衡收益，提高评估结果准确率。

经过计算 $\sum_{l=1}^{k}P_{ij}^l(Q_l\mid q_l,S_i^1,S_i^2)S_l$ 可知，其攻击行为对网络整个安全结构造成了一定影响，在攻击成功后的网络状态 q_l 熵差为：

$$\Delta Z=\sum_i I_i\Delta Z_{host_{ij}}+\sum_j\theta\Delta Z_{link_{ij}}$$
$$\Delta Z_{host_{ij}}=\Delta Z_m \tag{2}$$

公式（2）中：$\Delta Z_{host_{ij}}$ 为网络整个设备 $host_i$ 被攻击后的各个指标熵差相加后的总值，i 为设备数量，m 为可用性指标，I_i 为设备重要程度，$\Delta Z_{link_{ij}}$ 为攻击行为对链路影响；θ 为链路影响因子，可直观反映网络传输层受到影响程度。充分考虑防御者与攻击者双方利益，需进行模型对抗量化效果评比，通过评比效果可发现攻击行为对网络主动防御造成的损害情况，如果攻击行为大于防御行为，那么说明攻击者采用的攻击手段已经成功入侵了主动防御体系，查看设备安全使用情况。在既定安全性环境下使用 (S_i^1,S_i^2) 防御者收益结果作为量化评估标准。

1.4 安全性评估

将网络安全问题视为防御者和攻击者的多个阶段的博弈，每个阶段都对应一个安全状态，通过求解可对安全状态下的混合策略进行均衡处置，由此可获取网络安全状态下的双方对抗最优方案，引入网络熵展开定量分析，具体评估如下所示：

设置 in 为大数据驱动下主动防御网络的各个指标信息和双方对抗行为信息；out 为攻击行为的预测和对抗量化效果值。将初始化价值向量设为：

$$H^0=(h_1^0,h_2^0,\cdots,h_k^0)=(0,0,\cdots,0),$$

重复

$$for\quad each\quad Q_l\in q\quad do$$
$$for\quad each\quad q_{ij}^l\in \mathbf{Q}_l\quad do$$

用 h_l 代替模型公式（1）中的 Q_l

$$end\quad for$$

计算各个阶段的熵差，结合数学模型，更新价值向量，计算混合均衡得出 $(Q_l\mid q_l,S_i^1,S_i^2)S_l$ 和状态概率，分析主动防御网络安全情况，由此实现对主动防御网络安全性评估。

分析主动防御原理，构建评估数学模型，引入网络熵对抗量化技术来权衡收益，减少干扰因素影响，提高评估结果准确率，将防御者收益结果作为量化评估标准，实现对主动防御网络安全性评估。

2 实验

为了验证大数据驱动下主动防御网络安全性评估技术研究的合理性进行了如下实验，设计网络拓扑结构如图 4 所示。

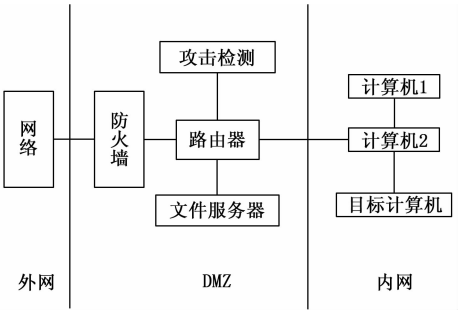

图 4 网络拓扑结构

攻击者首先需要获取权限才能进入攻击状态，为此需获取目标主机 root 的访问权限，由于防火墙只能允许主机对 Apache 服务器进行访问，对于内网服务器和主机并没有访问权力，受到限制，内网主机是不允许防火墙访问的。

如果攻击者获取了主机的访问权限，那么其内部的网络权限为最低初始访问权限，攻击者只能完成一次攻击，不能同时进行多个攻击行为。

2.1　实验条件设置

针对主动防御网络可用性，将吞吐量、传输延迟和故障情况作为指标，利用网络熵描述链路属性，在初始阶段，攻击者需获取想要攻击设备的权限才能进行攻击。针对 Apache 服务器攻击者需要利用 Smtp 服务所存在安全隐患发起攻击行为，由此获得用户使用权限，然后利用主机存在的安全漏洞获取 Root 权限。计算攻击行为直接受益情况，根据构建模型，使用对抗量化评估规则获取对抗最优方案。

2.2　实验结果与分析

将吞吐量、传输延迟和故障情况作为指标对主动防御网络安全性评估技术准确性进行实验验证。

2.2.1　吞吐量对技术准确性影响结果与分析

吞吐量是对网络在单位时间内成功传送数据的数量，也是说吞吐量是指在没有帧丢失情况下，设备能够接收并转发最大数据速率。将测试接入点选在链路两端以太网网络上，通过改变帧长度，在接收器上计算收到帧速，由于吞吐量测试是必须在线进行的，即不能中断现有网络业务和网络连接。将传统评估技术与引入网络熵的模型对抗量化评估技术在吞吐量不同情况下精准度对比结果与分析，如图 5 所示。

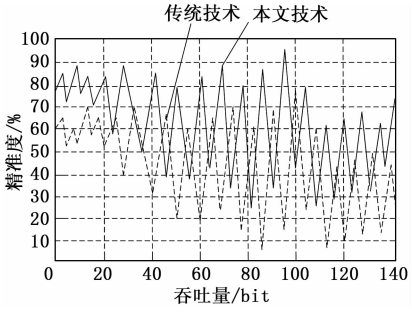

图 5　两种技术在不同吞吐量下精准度对比结果

由图 5 可知：实线为引入网络熵的模型对抗量化评估技术，虚线为传统评估技术。当吞吐量为 20 时，传统评估精准度达到最高为 70%，当吞吐量为 83 时，传统评估精准度达到最低为 5%；而当吞吐量为 95 时，引入网络熵的模型对抗量化评估技术达到最高为 96%，当吞吐量为 81 时，引入网络熵的模型对抗量化评估技术达到最低为 25%。由此可知，吞吐量对评估技术准确性具有严重影响，尤其是对传统技术影响极大，与对抗量化评估技术精准度最高值相差 26%，最低值相差 20%。吞吐量对引入网络熵的模型对抗量化评估技术影响效果较小。

2.2.2　传输延迟对技术准确性影响结果与分析

传输延迟是发送接收处理时间、电信号响应时间、介质中传输时间三个时间的总和，使用 Intel Core i7 处理器对内存延迟进行深度测试，将传统评估技术与引入网络熵的模型对抗量化评估技术在传输出现延迟情况下，评估精准度对比结果与分析，如表 3 所示。

表 3　两种技术在传输延迟下精准度对比结果

实验次数	传统评估技术				对抗量化评估技术			
	传输延迟			精准度/%	传输延迟			精准度/%
	发送接收	信号响应	介质传输		发送接收	信号响应	介质传输	
1	—	正常	正常	68	—	正常	正常	78
2	正常	—	正常	75	正常	—	正常	93
3	正常	正常	—	69	正常	正常	—	89
4	—	—	正常	45	—	—	正常	69
5	—	正常	—	58	—	正常	—	61
6	正常	—	—	37	正常	—	—	52

表 3 中，“—”代表出现延迟，对比两种评估技术发现，当发送接收处理出现延迟时精准度影响最大，传统技术比对抗量化技术评估精准度要低 10%，电信号响应出现延迟时精准度影响最小，对抗量化技术比传统技术评估精准度要高 18%。由此可知，传输延迟对传统评估技术影响效果较大，使用对抗量化评估技术精准度较高。

2.2.3　故障出现对技术准确性影响结果与分析

出现故障原因有许多种，包括网络适配器（网卡）设置与计算机资源有冲突、网吧局域网中有两个网段，其中一个网网段的所有计算机都不能上因特网等。针对故障出现对技术准确性影响需将传统技术与对抗量化评估技术进行对比，结果如图 6 所示。

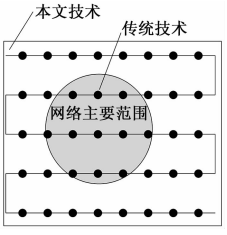

图 6　两种技术在故障点出现情况下对比结果

由图 6 可知：灰色部分为网络主要工作范围，白色部分为超出主要工作范围外的网络覆盖区域。采用传统评估技术只能对网络主要范围内的故障点进行处理，无法对超出该范围的故障点处理，导致评估精准度下降，无法准确获取主动防御网络的安全风险；而采用引入网络熵对抗量化评估技术不仅可以对灰色范围内故障点处理，还可以对该范围外的网络覆盖区域进行处理，不会对精准度有较大影响。

2.3　实验结论

根据上述对比的实验结果，可得出实验结论：大数据驱动下主动防御网络安全性评估技术研究具有合理性。

分别将吞吐量、传输延迟和故障情况作为指标进行实验验证，可知吞吐量对评估技术准确性影响最大，传统技术比对抗量化评估技术精准度最高值低 26%，而最低值相差 20%；传输延迟对传统评估技术影响效果较小，使用对抗量化评估技术精准度较高；在网络主要范围内出现故障

对于对抗量化技术来说，并不会影响评估精准度。

3 结束语

在大数据驱动下，引入网络熵对抗量化技术可将主动防御网络的安全性问题转化为多个阶段动态分析问题，为管理者提供有效评估措施，也为网络安全加固提供重要决策。由于大数据技术快速发展，促使引入网络熵模型对抗量化评估技术进入危险潜藏期，针对定期引入安全方案还无法实施，对于动态化大数据应用中心安全评估还有待考察。

参考文献：

[1] 陈 臣. 基于大数据驱动的图书馆动态网络性能评估和服务质量保证研究 [J]. 图书馆理论与实践，2016，11 (8)：89－93.

[2] 王 锋，王翔宇，秦文臻. 大数据驱动的高等教育质量监测评估关键技术研究 [J]. 黑龙江高教研究，2017，10 (6)：80－83.

[3] 孙远芳，段翠华，张培颖. 大数据驱动的未来网络：体系架构与应用场景 [J]. 中国电子科学研究院学报，2017，12 (5)：463－468.

[4] 赵大伟. 大数据技术驱动下的互联网消费金融研究 [J]. 金融与经济，2017，26 (1)：41－45.

[5] 汪 磊，许 鹿，汪 霞. 大数据驱动下精准扶贫运行机制的耦合性分析及其机制创新——基于贵州、甘肃的案例 [J]. 公共管理学报，2017，25 (3)：135－143.

[6] 李 信，李旭晖，陆 伟，等. 大数据驱动下的图书情报学科热点领域挖掘——面向 WOS 题录数据的实证视角 [J]. 图书馆论坛，2017，37 (4)：49－57.

[7] 计国君，余木红，KimHuaTan. 大数据驱动下的全渠道供应链服务创新决策框架 [J]. 商业研究，2016，62 (8)：152－162.

[8] 祝 丹，陈立双. 大数据驱动下统计学人才培养模式研究 [J]. 统计与信息论坛，2016，31 (12)：87－92.

[9] 尹 浩，乔 波. 大数据驱动的网络信息平面 [J]. 计算机学报，2016，30 (1)：126－139.

[10] 刘钊远. 大容量数字交换芯片 MT90820 及其应用 [J]. 电子设计工程，2017，20 (1)：8－12.

(上接第 275 页)

图 8 系统运行结果在 CyVOD 首页上的部分展示

模式分析网页变化的特征，用表征科技类信息的链出链接的变化作为网页变化的依据，结合时间感知相似性协方差矩阵和最大相似度阈值最大精度的优化页面的爬行计划。以网页最佳爬行时间戳序列作为网页刷新策略来增量式地更新分布式并行爬虫。实验证明，相比定期频繁的刷新策略，该方法能以较小的刷新代价获得较好的爬虫性能和更新质量。

参考文献：

[1] 周德懋，李舟军. 高性能网络爬虫：研究综述 [J]. 计算机科学，2009，36 (8)：26－29.

[2] 徐雁飞，刘 渊，吴文鹏. 社交网络数据采集技术研究与应用 [J]. 计算机科学，2017，44 (1)：277－282.

[3] Guo R，WangH Z，Chen M W，et al. Parallelizing the extraction of fresh information from online social networks [J]. Future Generation Computer Systems，2016，59：33－46.

[4] Xia J，Wan W G，Liu R Z，et al. Distributed web crawling：A framework for crawling of micro－Blog data [A]. International Conference on Smart & Sustainable City & Big Data [C]. Shanghai，China：IET Conference Publications，2016：62－68.

[5] Su L P，Wang F X. Web Crawler Model of Fetching Data Speedily Based on Hadoop Distributed System [A]. Proceeding of2016 IEEE 7th International Conference on Software Engineering and Service Science [C]. Beijing：IEEE，2016：927－931.

[6] Huang Q，Li Q，Yan Z，et al. A novel incremental parallel web crawler based on focused crawling [J]. Journal of Computational Information Systems，2013，9 (6)：2461－2469.

[7] 周中华，张惠然，谢 江. 基于 Python 的新浪微博数据爬虫 [J]. 计算机应用，2014，34 (11)：3131－3134.

[8] 黄志敏，曾学文，陈 君. 一种基于 Kademlia 的全分布式爬虫集群方法 [J]. 计算机科学，2014，41 (3)：124－128.

[9] Gu R，Jiang J F. Research and implementation of topic crawler based on Hadoop [J]. Applied Mechanics & Materials，2014，651－653：1896－1900.

[10] Tan Q，Zhuang Z，Mitra P，et al. A clustering－based sampling approach for refreshing search engine's database [A]. Tenth International Workshop on the Web and Databases [C]. Beijing，2007：1－6.

[11] Calzarossa M C，Tessera D. Modeling and predicting temporal patterns of web content changes [J]. Journal of Network & Computer Applications，2015，56：115－123.

[12] Olston C，Pandey S. Recrawl scheduling based on information longevity [A]. Proceeding of the 17th International Conference on World Wide Web 2008 [C]. Association for Computing Machinery，2008：437－446.

[13] 刘 慧. 基于增量爬虫与微博的视频资源推广技术研究 [D]. 武汉：华中科技大学，2012.

[14] Lu F，Tang Z Y，Liao X F，et al. An incremental crawler for web video based on content longevity [A]. Proceedings－2013 8th Annual China Grid Conference [C]. IEEE Computer Society，2013：98－102.

[15] Gupta K，Mittal V，Bishnoi B，et al. AcT：Accuracy－aware crawling techniques for cloud－crawler [J]. World Wide Web－Internet & Web Information Systems，2016，19 (1)：69－88.