第35卷 第24期 计 算 机 工 程 2009年12月
Vol.35 No.24 Computer Engineering December 2009
·博士论文· 文章编号：1000—3428(2009)24—0033—03 文献标识码：A 中图分类号：TP393

Web使用挖掘在信任管理中的应用

赵 洁[1,2]，肖南峰[1]，钟军锐[3]
(1. 华南理工大学计算机科学与工程学院，广州 510640；2. 广东工业大学管理学院，广州 510520；
3. 暨南大学附属第一医院计算机中心，广州 510630)

摘 要：通过研究IIS的工作原理、.Net程序架构及其宿主环境，在HTTP管道中拦截用户请求，创建一种新的Web日志，以聚类等方法对新日志进行挖掘，将信息数据应用于电子商务系统的信任管理中。实验证明，基于新日志的信息可以提高服务器的各项性能，约束用户的商业行为。

关键词：信任管理；Web使用挖掘；Web日志；IIS插件

Application of Web Usage Mining in Trust Management

ZHAO Jie[1,2], XIAO Nan-feng[1], ZHONG Jun-rui[3]
(1. School of Computer Science & Engineering, South China University of Technology, Guangzhou 510640;
2. School of Management, Guangdong University of Technology, Guangzhou 510520;
3. Computer Center, The First Affiliated Hospital of Jinan University, Guangzhou 510630)

【Abstract】 By studying IIS, .Net framework and its host environment, user requests are held up to create a kind of new Web log. It uses clustering and other methods to mine new log, and applies information data to trust management of E-commerce. Experiments show that information from new log can enhance all kinds of performances of server and restrict user's trade behaviors.

【Key words】 trust management; Web usage mining; Web log; IIS plug-in

在Web使用挖掘中，服务器软件自动记录的日志目前最为常用[1]，但会产生许多不必要信息。预处理是保证Web使用挖掘质量的关键，文献[2]提出多种预处理方法，但难度较大，效果也难以令人满意。如今电子商务面临严峻的网络安全问题。传统的安全技术和手段，尤其是安全授权机制，不再适用于解决Web安全问题[3]。信任管理为解决电子商务、分布式应用等系统的安全问题提供了新思路。本文创建了一种基于IIS的新型日志，选择性地记录有用信息，可简化预处理工作，建立用户历史行为证据，提供量化统计数据，通过聚类等算法分析，将数据应用于电子商务的信任管理。

1 基于IIS插件Web日志的设计及实现

1.1 模式设计

参考W3C的Web日志，所设计的逻辑模式如表1和表2所示，篇幅有限，表1仅列出部分字段。

表1 行为日志的数据库模式

字段代号	字段名称
RequestID	请求编号
UserUnid	用户全局标识
RawUrl	访问路径
FilePath	文件路径
InfoPath	附加路径
csBytes	接收字节数
scBytes	发送字节数
TimeProcessTaken	处理时间
UserAgent	用户代理
Referer	页面引用
DateTime	请求开始时间
Status	访问状态
IsExceedAuthority	尝试越权访问

表2 商业日志的数据库逻辑模式

字段代号	字段名称
OrderID	订单号
StateID	状态号
StateDesc	状态描述
DateTime	时间
UserUnid	用户全局标识

字段说明如下：

UserUnid为用户表主键，在此作为外键，建立多表关系。匿名用户统一标志为00000000-0000-0000-0000-000000000000。

RawUrl, FilePath, InfoPath记录用户访问的精确路径信息，粒度可细化至页面按钮级。

其他用于信任管理控制统计的字段在下文介绍。

1.2 日志的实现

Asp.Net程序处理用户请求流程如图1所示[4]，其中，拦截用户请求需在.Net Framework底层进行，捕捉请求的3个类：TBO(TrustBusinessObject), CTBO(ConnectionTrustBO)和BTBO(BehaviorTrustBO)需要实现IHttpModule接口和IRequiresSessionState接口，用于处理HttpModule上的事件，

基金项目：国家自然科学基金委员会与中国民用航空总局联合基金资助项目(60776816)；广东省自然科学基金资助重点项目(8251064101000005)；广东省科技计划基金资助项目(2007B060401007)；广东工业大学青年基金资助项目(072058)

作者简介：赵 洁(1979－)，女，博士研究生，主研方向：智能计算，电子商务；肖南峰，教授、博士、博士生导师；钟军锐，学士

收稿日期：2009-04-20 **E-mail：**kitten-zj@163.com

通过 Web.config 设置类之间的优先级。

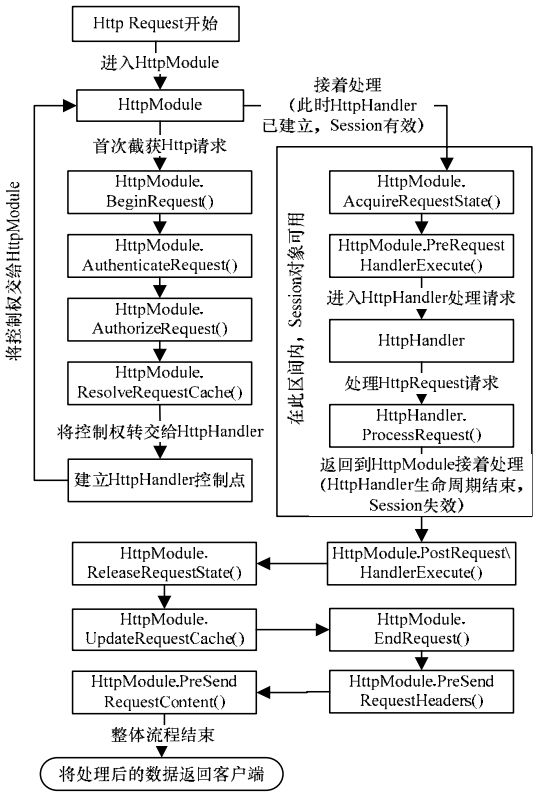

图 1　Asp.Net 处理用户请求的流程

工作流程具体如下：

(1)处理 BeginRequest 事件。TBO 首先处理该事件，实例化一个 TraceStream 对象，将请求信息装载到该对象包含的 TMRequest 对象中。CTBO 将 TMRequest 对象传递给处理模块 1 验证 IP 合法性，通过则执行后续处理；否则，终止本次会话，标记为越权访问，记入用户行为日志中。

(2)处理 AcquireRequestState 事件。TBO 从 Session 中获得用户全局标识存储到 TMRequest 对象中。CTBO 再次将 TMRequest 对象传递给模块 1，验证请求资源所需权限。验证后的处理与 IP 合法性验证方法相同。BTBO 将 TRequest 对象传递给处理模块 2 进行预测和控制决策。若模块 2 不接收该请求，终止本次会话；否则，执行 HttpModule 后续动作。

(3)处理 EndRequest 事件。BTBO 将 TraceStream 对象收集的信息装载到 TMRequest 对象中，传递给模块 2，模块 2 将本次会话记录添加到用户行为日志中，评价信任等级并更新数据库，形成历史经验。

1.3　日志特点

本文日志与服务器自动产生的日志相比，有如下特点：(1)具伸缩性：属性可根据实际应用添加或减少，数据是定制的、经处理的信息。(2)简化清理数据工作：自动过滤无用记录，不记录静态元素，如图片、音频、HTML。(3)易于识别用户：明确区分出注册用户和匿名用户，启发规则可充分利用 IP+代理信息区分匿名用户。(4)易于补充路径：页面“回退”时，当页面含有 JS 和 IFrame，则访问服务器，需记入日志，其他情况均不记录。(5)提供精确统计数据：RawUrl 等记录的信息可精细至页面功能，冗余信息快捷实现数据统计。

2　日志在信任管理中的应用

2.1　总体设计

设计使用分布式工作环境，信任管理以可代理、可配置形式应用到电子商务中，如图 2 所示，具体步骤如下[5]：

(1)请求服务。

(2)提交用户身份及请求资源信息，验证身份与权限。

(3)访问。

(4)返回身份验证信息和资源访问权限。

(5)返回验证结果。

(6)如 SP 拒绝，转(13)；否则，提交 Agent1 的管理请求。

(7)访问。

(8)返回以往用户行为信任的统计数据。

(9)对用户行为进行预测，得出信任等级并进行控制决策，返回结果。

(10)如接收，则访问 DB1；否则，跳(13)。

(11)返回资源。

(12)返回资源。

(13)提交本次交互的用户行为信息。

(14)添加用户的行为信息到用 DB3。

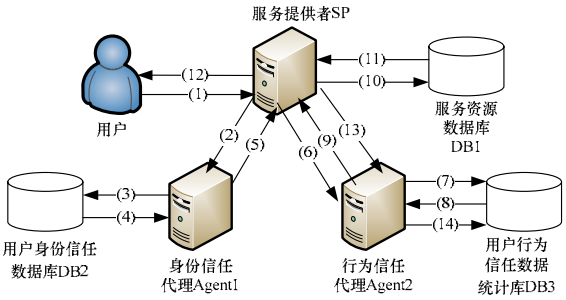

图 2　行为信任预测与控制整体设计

第 1 节介绍的行为日志存储于 DB2 中。实现时需要将 1.2 节中的模块 1 替换成身份行为代理，模块 2 替换成行为信任代理。

2.2　信任属性判断依据及等级描述

用户信任预测使用贝叶斯网络模型[6]，具体方法已另作讨论，此处不再重复。信任等级 T 及分解后的信任属性如下：(1)安全信任属性：证据为日志中的 IsExceedAuthority 属性。(2)可靠性信任属性：证据为日志中的 Status 属性。当 Status=200，表示会话正常完成。(3)性能信任属性：证据为日志中的 Cs Bytes, Sc Bytes 及 Time Process Taken。(4)商业信任属性：证据为商业行为日志中的订单完成率。

信任等级划分为：A(非常信任)，B(信任)，C(比较信任)，D(基本信任)，E(不信任)，由信任代理维护。根据上述等级，信任需求等级分为：E(全部接受)，D(需要较低信任)，C(需要一般信任)，B(需要较高信任)，A(需要非常高信任)，S(全部拒绝)。信任等级评价模式如表 3 所示。

表 3　信任等级评价记录

字段代号	字段名称
ExperienceID	经验编号
ClientIP	客户端 IP 地址
UserUnid	用户全局标识
DateTime	时间
TrustLevel	信任等级
SecurityLevel	安全等级
ReliabilityLevel	可靠性等级
PerformanceLevel	性能等级
BusinessLevel	商业等级

2.3 信任区间的确定

本文假定所有属性信任等级服从正态分布，采用聚类算法确定区间的分布，举例如下。

设一段时间内的日志有 11 779 条，挖掘模型中使用 Cs Bytes, Sc Bytes, Time Process Taken 这 3 个属性，聚类后 5 个类别的数量分别为 1 381, 2 811, 3 622, 2 523, 1 442，得到性能分布区间为[0,0.12], (0.12,0,36], (0,36,0.67], (0,67,0,88], (0.88,1]。可每隔一段时间进行分析，以调整贝叶斯算法的参数。

2.4 信任等级评定

评定信任等级需要日志中的多项数据：

(1)安全信任等级 S。CE_x 表示用户 x 尝试越权访问次数，f_x 表示越权访问频率，则有：$f_x=\frac{CE_x}{n}$。$\overline{f}$ 表示所有用户的平均越权访问频率，$D(f)$ 表示其方差，等级符合 $f\sim N\left(\overline{f},D(f)\right)$ 正态分布。

(2)可靠性信任等级 R。计算方法同上。

(3)性能信任等级 P。证据包括用户发送字节数 cs，服务器发送字节数 sc，请求处理时间 pt。引入一个描述性能的参数 Dp，定义为

$$Dp=\omega_1\times cs+\omega_2\times sc+\omega_3\times pt \tag{1}$$

其中，$\omega_i\left(i=1,2,3\right)$ 为影响因子，由专家或历史统计数据得出，计算方法同上。

(4)商业信任等级 B。证据是用户订单完成率，越低则用户信任等级越低。计算方法同上。

(5)总体信任等级 T。用 j 表示安全信任等级、k 表示可靠性信任等级、m 表示性能信任等级、n 表示商业信任等级，总体信任等级 T 的计算方法如下：

$$T=j\times w_s+k\times w_r+n\times w_B \tag{2}$$

其中，$w_S,w_R,w_P,w_B\geq 0$ 且 $w_S+w_R+w_P+w_B=1$，w_S,w_R,w_P,w_B 分别表示各属性等级的权重，由专家给出或统计数字得出。

3 实例分析

3.1 参数说明

(1)式(1)中的 3 个权重为 1, 3, 4，式(2)中参数为 0.3, 0.2, 0.25, 0.25。

(2)各信任等级分布区间如下：安全信任属性：[0, 0.13], (0.13, 0.34], (0.34, 0.61], (0.61, 0.85], (0.85, 1]；可靠性信任属性：[0, 0.05], (0.05, 0.20], (0.20, 0.60], (0.60, 0.85], (0.85, 1]；性能信任属性：[0, 0.20], (0.20, 0.30], (0.30, 0.75], (0.75, 0.85], (0.85, 1]；商业信任属性：[0, 0.25], (0.25, 0.35], (0.35, 0.70], (0.70, 0.88], (0.88, 1]。

(3)所有用户的平均越权访问频率为 10%，方差为 5%；性能平均值为 8 000，方差为 10 000 000；平均可靠性为 75%，方差为 4%；商业属性的订单完成率为 70%，方差为 5%。

(4)用户 A 的历史评价记录如表 4 所示，尝试越权率为 20%，订单完成率为 85%。

表 4 用户 A 信任等级的历史评价记录

总体信任等级	安全信任等级	可靠性信任等级	性能信任等级	商业信任等级
4	2	1	5	4
4	2	2	4	4
3	2	3	3	4
3	2	4	2	4
3	2	5	1	4

(5)对无历史数据的新用户，以比较信任的等级处理。

根据 A 的尝试越权率算得安全信任等级为 2，商业信任等级为 4。根据 $\max(p(T_i\mid S_2,B_4))=\max(\frac{\left|S_2\cap B_4\cap T_i\right|}{\left|S_2\cap B_4\right|})$ 预测用户的总体信任等级为 3。

3.2 用户行为评价

设本次会话 cs=100, sc=2 000, pt=50，算得性能参数 Dp=6 300，性能信任等级为 4。假设用户在过去 99 次会话中有一次非正常终止，本次会话正常结束，则会话完成率为 99%，可靠性信任为 5。本次会话结束后，用户的总体信任等级为 0.3×2+0.2×5+0.25×4+0.25×4=3.6，总体信任等级为 4，其评价记录的更新情况如表 5 所示。

表 5 用户 A 信任等级的历史评价记录(更新后)

总体信任等级	安全信任等级	可靠性信任等级	性能信任等级	商业信任等级
4	2	1	5	4
4	2	2	4	4
3	2	3	3	4
3	2	4	2	4
3	2	5	1	4
4	2	5	4	4

4 应用效果分析

取 6 周的日志数据用聚类算法分析区间变化，见图 3 和图 4。不信任等级性能区间有较明显的下降趋势，信任等级性能区间显式上扬，说明服务器的性能得到提高，五级区间趋于正态分布。商业信任区间变化类似。

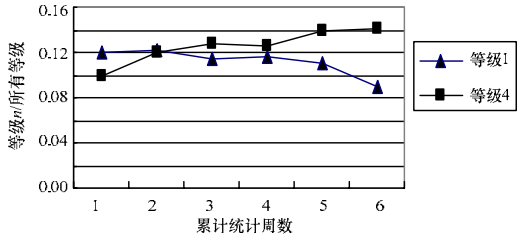

图 3 性能信任等级分布区间变化

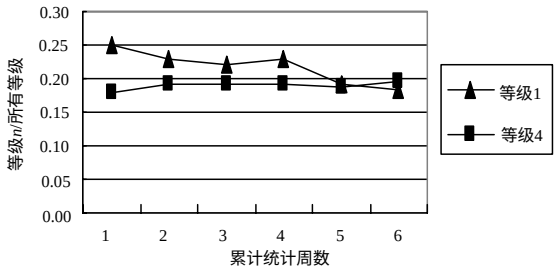

图 4 商业信任等级区间变化

5 结束语

本文深入研究了 IIS 和.Net 底层架构，设计并实现了基于 IIS 插件的 Web 日志，与目前常用的 Web 日志相比，该日志免除了大量的预处理工作，在用户识别、会话识别、路径补充等方面更具易操作性和精确度。通过聚类等方法对日志进行挖掘，并应用于信任管理中。实验验证本日志可为行为信任的预测和控制提供实用数据，通过预测和控制，增强了服务器的安全性、可靠性，并有效约束了用户的商业行为。下一阶段将对此 Web 日志进行更深入的分析和研究。

(下转第 38 页)