

(19) 中华人民共和国国家知识产权局



(12) 发明专利申请

(10) 申请公布号 CN 102664935 A

(43) 申请公布日 2012.09.12

(21) 申请号 201210100189.2

(22) 申请日 2012.04.06

(71) 申请人 北京锐安科技有限公司

地址 100044 北京市海淀区中关村南大街乙
56 号方圆大厦 9 层

(72)发明人 李金伟

(74) 专利代理机构 北京君尚知识产权代理事务

所（普通合伙） 11200

代理人 余功勋

(51) Int. Cl.

H04L 29/08 (2006, 01)

G06E 17/30 (2006-01)

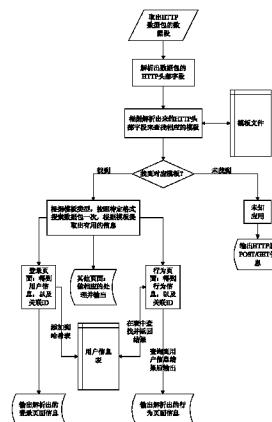
权利要求书 2 页 说明书 5 页 附图 1 页

(54) 发明名称

一种 WEB 类用户行为和用户信息的关联输出方法及系统

(57) 摘要

本发明公开了一种WEB类用户行为和用户信息的关联输出方法及系统，属于数据通信领域。本方法为：1) 建立网站的登录页面模板文件和行为页面模板文件，并在服务器内存中形成一模板链表；2) 用户登录该网站时，解析获取数据包头部信息，并在所述模板链表中查找匹配的模板文件，还原出相应的页面信息；3) 根据还原出的ID，对还原出的获取用户信息和行为信息进行关联后输出。本系统包括模板导入模块，用于读取模板并解析成模板链表操作；模板查找与内容提取模块，用于解析数据包头与对应的模板进行匹配，然后利用匹配模板对相应的页面进行解析；用户信息模块，用于维护用户信息哈希表。本发明可高效、便捷的将用户行为和用户信息关联输出。



1. 一种 WEB 类用户行为和用户信息的关联输出方法,其步骤为 :

1) 根据同一网站用户登录页面数据包和用户行为页面数据包,建立该网站的登录页面模板文件和行为页面模板文件并保存到服务器中;所述服务器中设有一用户信息哈希表、一WEB 类应用内容还原系统;

2) WEB 类应用内容还原系统读取所述登录页面模板文件和行为页面模板文件,并在该服务器内存中形成一模板链表;

3) 当用户登录该网站时,WEB 类应用内容还原系统解析获取的 HTTP 数据包头部信息,并在所述模板链表中查找匹配的模板文件;

4) 当匹配模板文件为登录页面模板文件时,则利用该登录页面模板文件解析当前 HTTP 数据包获取用户信息和 ID,并在所述用户信息哈希表中添加一个用户信息节点;

5) 当匹配模板文件为行为页面模板文件时,则利用该行为页面模板文件解析当前 HTTP 数据包,还原出行为信息和 ID;并在所述用户信息哈希表中查找与当前 ID 匹配的节点,得到当前用户操作行为对应的用户信息;

6) 将查找到的用户信息连同用户操作行为内容一并按照设定格式输出。

2. 如权利要求 1 所述的方法,其特征在于所述步骤 1) 中,首先提取所述用户登录页面数据包和用户行为页面数据包中的 WEB 类应用数据包特征,然后按照 WEB 类描述语言语法建立所述登录页面模板文件和行为页面模板文件。

3. 如权利要求 2 所述的方法,其特征在于所述 WEB 类应用数据包特征为:HTTP 头部字段、网站登陆信息、发帖内容。

4. 如权利要求 1 所述的方法,其特征在于所述 WEB 类应用内容还原系统按照 WEB 类描述语言语法解析所读取的每个模板文件,把解析出的关键字组成一个结构体,然后形成所述模板链表。

5. 如权利要求 1 所述的方法,其特征在于所述用户信息包括用户名和用户密码;所述用户信息节点包括客户端 IP、用户信息和 ID;所述用户信息哈希表根据用户信息节点的客户端 IP 建立索引。

6. 如权利要求 5 所述的方法,其特征在于以客户端 IP 作为索引值,在所述用户信息哈希表中查找与当前 ID 匹配的节点。

7. 一种 WEB 类用户行为和用户信息的关联输出系统,其特征在于包括模板导入模块、模板查找与内容提取模块、用户信息维护模块、输出模块;其中,所述模板导入模块包括模板文件,读取模板并解析成模板链表操作;所述模板查找与内容提取模块包括识别出 HTTP 协议,并解析 HTTP 头部分字段,与对应的模板进行匹配,然后利用匹配模板对相应的页面进行解析;所述用户信息模块,用于维护用户信息哈希表;所述输出模块,用于对关联上的用户信息进行输出。

8. 如权利要求 7 所述的关联输出系统,其特征在于所述模板文件包括登录页面模板文件和行为页面模板文件;根据同一网站用户登录页面数据包和用户行为页面数据包中的 WEB 类应用数据包特征,按照 WEB 类描述语言语法建立所述登录页面模板文件和行为页面模板文件。

9. 如权利要求 7 或 8 所述的关联输出系统,其特征在于所述模板导入模块按照 WEB 类描述语言语法解析所读取的每个模板文件,把解析出的关键字组成一个结构体,形成一模

板链表。

一种 WEB 类用户行为和用户信息的关联输出方法及系统

技术领域

[0001] 本发明属于数据通信领域,涉及一种 WEB 类用户行为和用户信息的关联输出方法及系统。

[0002] 背景资料

[0003] WEB 类应用基于 HTTP 协议,种类繁多,包括目前比较热门的社交类网站、邮箱类网站、BBS 类网站、博客类网站、赌博类网站、搜索类网站、在线聊天类网站等。还原 WEB 类应用的登陆与内容,对网络安全、色情监控等有着非常重要的作用。

[0004] 本发明利用一个高效的 web 类应用内容还原系统(一种还原 WEB 类应用内容的方法及其系统,申请号 201010612835.4),该系统通过一定规则生成 WDSL 文件(本文中统称“模板”),模板中定义出关键字,然后在原始数据包中高速寻找模板中对应的关键字,从而得到还原页面内容的目的。由于本文的重点不在于此系统,所以此处不再赘述。根据此系统,可以快速有效的还原出一个页面中需要的内容。

[0005] 网民登录的页面和发帖、发邮件等操作的页面都可以经过该系统处理还原出来,但是这两类页面却很难关联到一块。行为页面在数据经过还原后一般没有发帖人的相关信息,从数据包中很难还原出此操作的执行者信息,而没有操作者的操作信息将会大打折扣,所以找到一种简单的办法把用户登录信息和操作信息关联到一起,就成为数据还原非常重要的一个步骤。

[0006] 目前大部分网站采用单点登录方式对用户进行认证,当用户第一次访问应用系统的时候,因为还没有登录,会被引导到认证系统中进行登录;根据用户提供的登录信息,认证系统进行身份效验,如果通过效验,应该返回给用户一个认证的凭据--ticket;用户再访问别的应用的时候,就会将这个 ticket 带上,作为自己认证的凭据,应用系统接受到请求之后会把 ticket 送到认证系统进行效验,检查 ticket 的合法性。

发明内容

[0007] 针对目前对 WEB 类应用协议内容还原技术的局限性,本发明的目的在于提供一种 WEB 类用户行为和用户信息的关联输出方法及系统;本发明通过利用 ticket 的特性,在数据还原时,先把用户登录页面和用户行为页面还原出来,然后根据 ticket 来把用户登录页面和用户行为页面进行关联。

[0008] 本发明利用网站对登录信息的记录和保存,提出了一种基于模板以及关联 ID 的方法来还原 WEB 类应用内容,并关联到用户信息的方法。

[0009] 本发明利用一个高速的 web 类应用内容还原系统,该系统通过一定规则生成模板,模板中定义出关键字,然后在原始数据包中高速寻找模板中对应的关键字,从而得到还原页面内容的目的。根据此系统,可以快速有效的还原出一个页面中需要的内容。

[0010] 本发明是在该还原系统的基础上,对两个甚至多个页面之间的信息进行关联,从而达到获得操作行为中用户信息的目的。为方便描述,本文中在用户信息出现的页面,比如登录、注册等统一称为“登录页面”,对应的模板成为“登录模板”;在用户进行发帖、发邮件、

发博客等操作统一称为“行为页面”，对应的模板称为“行为模板”。行为页面需要关联到对应的登录页面，才能获得对应的用户信息。

[0011] 本发明的技术方案为：

[0012] 一种 WEB 类用户行为和用户信息的关联输出方法，其步骤为：

[0013] 1) 根据同一网站用户登录页面数据包和用户行为页面数据包，建立该网站的登录页面模板文件和行为页面模板文件并保存到服务器中；所述服务器中设有一用户信息哈希表、一 WEB 类应用内容还原系统；

[0014] 2) WEB 类应用内容还原系统读取所述登录页面模板文件和行为页面模板文件，并在该服务器内存中形成一模板链表；

[0015] 3) 当用户登录该网站时，WEB 类应用内容还原系统解析获取的 HTTP 数据包头部信息，并在所述模板链表中查找匹配的模板文件；

[0016] 4) 当匹配模板文件为登录页面模板文件时，则利用该登录页面模板文件解析当前 HTTP 数据包获取用户信息和 ID，并在所述用户信息哈希表中添加一个用户信息节点；

[0017] 5) 当匹配模板文件为行为页面模板文件时，则利用该行为页面模板文件解析当前 HTTP 数据包，还原出行为信息和 ID；并在所述用户信息哈希表中查找与当前 ID 匹配的节点，得到当前用户操作行为对应的用户信息；

[0018] 6) 将查找到的用户信息连同用户操作行为内容一并按照设定格式输出。

[0019] 进一步的，所述步骤 1) 中，首先提取所述用户登录页面数据包和用户行为页面数据包中的 WEB 类应用数据包特征，然后按照 WEB 类描述语言语法建立所述登录页面模板文件和行为页面模板文件。

[0020] 进一步的，所述 WEB 类应用数据包特征为：HTTP 头部字段、网站登陆信息、发帖内容。

[0021] 进一步的，所述 WEB 类应用内容还原系统按照 WEB 类描述语言语法解析所读取的每个模板文件，把解析出的关键字组成一个结构体，然后形成所述模板链表。

[0022] 进一步的，所述用户信息包括用户名和用户密码；所述用户信息节点包括客户端 IP、用户信息和 ID；所述用户信息哈希表根据用户信息节点的客户端 IP 建立索引。

[0023] 进一步的，以客户端 IP 作为索引值，在所述用户信息哈希表中查找与当前 ID 匹配的节点。

[0024] 一种 WEB 类用户行为和用户信息的关联输出系统，其特征在于包括模板导入模块、模板查找与内容提取模块、用户信息维护模块、输出模块；其中，所述模板导入模块包括模板文件，读取模板并解析成模板链表操作；所述模板查找与内容提取模块包括识别出 HTTP 协议，并解析 HTTP 头部分字段，与对应的模板进行匹配，然后利用匹配模板对相应的页面进行解析；所述用户信息模块，用于维护用户信息哈希表；所述输出模块，用于对关联上的用户信息进行输出。

[0025] 进一步的，所述模板文件包括登录页面模板文件和行为页面模板文件；根据同一网站用户登录页面数据包和用户行为页面数据包中的 WEB 类应用数据包特征，按照 WEB 类描述语言语法建立所述登录页面模板文件和行为页面模板文件。

[0026] 进一步的，所述模板导入模块按照 WEB 类描述语言语法解析所读取的每个模板文件，把解析出的关键字组成一个结构体，形成一模板链表。

[0027] 本发明的主要技术内容为：

[0028] 1) 在实验环境下完成一次完整的网站登录和用户行为操作并分析数据包，首先获取该用户的行为页面和登录页面，然后根据 WEB 类描述语言语法完成登录页面和行为页面的模板，特别要注意的是需要在登录和用户行为操作数据包中找到两者之间的一个相同标识的 ID，此 ID 一般存在于 cookie 或者 url 中，确保通过登录模板和行为模板都能够取得此 ID，用此 ID 来关联登录页面和行为页面。模板文件是根据 WEB 类应用数据包特征，按 WEB 类描述语言语法，写成 WEB 类应用模板；针对某个具体网站所建立的模板保存在 web 应用还原系统中预先设定的位置。

[0029] 2) 系统读取 WDSL 格式的模板文件，并在内存中按照 WEB 类描述语言语法形成模板链表，然后系统处理所有经过系统的 HTTP 数据包，通过解析 HTTP 数据包头部信息，比如 POST/GET、HOST、URL 等信息，在模板链表中查找到与解析出的头部信息匹配的模板信息，从而通过模板还原出相应的页面信息。

[0030] 3) 系统中维护了一个用户信息哈希表，在登录页面匹配上模板时，系统通过模板获取用户名、密码等用户信息，并在用户信息哈希表中添加一个用户信息节点，节点包括了客户端 IP、用户信息和 ID，此节点以客户端 IP 为索引。

[0031] 4) 在行为页面匹配上模板时，按照模板规则（模板规则在 web 类应用还原系统中设定）还原出行为信息并解析出关联 ID，并以客户端 IP 作为索引值，在用户信息哈希表中查找和关联 ID 符合的节点，若找到，则可找到此次用户操作行为对应的用户信息。

[0032] 5) 根据查找到的用户信息，连同用户操作行为内容一并格式化，然后输出。

[0033] 所述步骤 1) 模板文件是按照 WEB 类应用数据包分析出特征字段，然后按照一定的 WEB 类描述语言语法写成模板文件，重点在于在登录页面模板和行为页面模板中需要能够取到一致的关联 ID。所述 WEB 类描述语言为根据 WEB 类数据包特征开发的描述语言；所述 WEB 类数据包特征为 HTTP 头部字段，网站登陆信息，发帖内容等。

[0034] 所述步骤 2) 系统按照模板解析单个页面信息，此过程虽复杂，但不是本专利研究的重点。

[0035] 所述步骤 3) 系统中维护一个用户信息哈希表，在解析用户登录页面后，把得到的用户信息连同关联 ID 一同添加到哈希表的节点中。所述用户信息哈希表为存储用户信息数据结构体在系统中组成的哈希链表，该哈希表中的节点保存了用户的账号、密码信息等，还包括了关联 ID，每个节点以客户端 IP 为索引值。

[0036] 所述步骤 4) 在出现行为页面后，解析出关联 ID，会利用客户端 IP 作为索引，查找用户信息表，从中查找和关联 ID 一致的节点，如果找到节点，则可以取出对应的用户信息。

[0037] 所述步骤 5) 根据查找到的用户信息，按照数据库的规则格式化数据，然后以文件的形式输出。

[0038] 一种 WEB 类用户行为和用户信息的关联输出系统，包括模板导入模块、模板查找与内容提取模块、用户信息维护模块，输出模块。所述模板导入模块包括模板文件，读取模板并解析成模板链表操作；所述模板查找与内容提取模块包括识别出 HTTP 协议，并解析 HTTP 头部分字段，与对应的模板进行匹配；所述用户信息模块，是对用户信息哈希表的维护；所述输出模块，是对关联上的用户信息进行输出的模块。具体为：

[0039] 所述模板导入模块主要负责将模板文件动态读入到系统，然后形成模板数据结构

链表,它首先扫描分类好的模板文件,接着把模板描述的关键字组成数据结构,然后形成链表状态。

[0040] 所述模板查找与内容提取模块是用来负责还原 WEB 类应用内容的。它先将 HTTP 数据包头分解,取出 POST/GET、HOST 和 URL 等字段值,然后用这些值进行模板匹配,如果有匹配模板,则可以进行模板关键字匹配,从而还原出应用协议内容。

[0041] 所述用户信息维护模块是负责维护一个用户信息哈希表,在登录页面和行为页面解析出来后,对这个用户信息表进行添加节点和查询节点等操作。

[0042] 所述输出模块是对页面解析出来的结果连同在用户信息链表中查找到的节点信息进行格式化,然后输出数据。

[0043] 与现有技术相比,本发明的积极效果为:

[0044] 1. 单纯从数据还原的角度来说,还原一个页面的技术并不难实现,但是目前并没有把两个相关的页面联系到一起,从而在现有不含用户身份信息的页面中得到用户信息,本发明在此技术上有创新性。

[0045] 2. 由于网站识别一个用户,往往是通过在 cookie 中记录相关信息,或者是记录一个通行的 ID,类似通行证,从而在该网站登录后就可以在网站的任何一个站点识别出该用户信息。如果用软件代码实现用户信息的关联,需要针对每一个网站进行代码实现,而本发明基于模板匹配机制,仅仅需要针对每个网站修改模板即可,相对代码实现更加的有效和便捷。

附图说明

[0046] 附图为本发明的方法流程图。

具体实施方式

[0047] 下面结合附图,进一步详细描述本发明的实施方案:

[0048] 在一个配有两块网卡的服务器上,比如配有两块网卡的高性能 RedFlag 服务器上,即可实施本系统,一块网卡作为数据源,另一块网卡用作数据通讯,模板文件可以存储在本机硬盘上,协议分析人员只需增加、更改模板文件,即可实现协议的增加与修正。输出文件可以输出到硬盘,也可以通过网络输出到其他系统。实施过程如附图所示,需要如下步骤:

[0049] (1)、增加模板文件

[0050] 分析人员需要先对所要分析的数据包进行分析,即需要捕获同一网站用户登录和用户行为页面的数据包,提取出关键内容特征,然后按照 WEB 类描述语言语法,写成相应协议模板。可以用 wireshark 等抓包工具获取数据包;针对不同网站,所建模板文件一般不同。

[0051] (2)、系统启动并导入模板

[0052] 系统启动后,各个模块初始化后开始正常运转,系统将模板文件读入,并按照 WEB 类描述语言语法解析每个模板文件,把解析出的关键字组成一个结构体,然后形成模板链表;通常在 web 类应用内容还原系统中定义有结构体,由于结构相对固定,用结构体可提高处理效率。

[0053] (3)、系统接收网卡数据包,并从应用层识别出 HTTP 数据连接

[0054] 系统从网卡上接收到数据后,在进行完传输层识别后,进行应用层协议识别。系统是根据 GET/POST 等 HTTP 头部标识字段进行 HTTP 协议识别,而不是根据端口,因为很多 HTTP 应用流量已经不通过 80 端口收发了,因为 HTTP 是基于 TCP 协议,所以只需有一个数据包被识别为 HTTP,整个 TCP 连接即可被识别为 HTTP 连接。

[0055] (4)、根据识别出的数据包进行模板匹配

[0056] 在识别出 HTTP 连接后,系统开始解析 HTTP 头部,并与模板链表里的模板结构进行比对,如果 HOST、URL 等关键字段匹配,则认为找到了合适的模板;如果未找到匹配模板,则输出解析出的 HTTP 头部信息。

[0057] (5)、根据匹配的模板从数据流中提取关键内容

[0058] 如果模板匹配成功即可提取出相关内容,从而还原出一个页面的关键内容。

[0059] (6)、根据匹配的模板类型关联用户信息

[0060] 根据匹配的模板类型进行相应处理,如果是匹配到登录模板,需要在还原出页面后把提取到的用户信息添加到用户信息哈希表中,如果是匹配到用户行为模板,则需要在哈希表中查找对应的用户信息。

[0061] (7)、将解析出的协议内容格式化输出

[0062] 输出模块的作用就是将协议内容格式化,然后输出。输出内容一般会存储到设定数据库,以便更好地分析利用,所以协议内容会按照数据库的格式进行格式化。

[0063] 到此,关联 WEB 类用户行为和用户信息实施完毕,如果想增加支持的 web 协议种类,只需增加模板即可实现。

