基于代理的 Web 页访问语义过滤与内容重现

周文刚,马占欣
(周口师范学院 计算机科学系,河南 周口 466001)

摘 要:对 Web 页进行必要的、有效的内容过滤对于营造健康、安全的网络环境具有重要的意义。重现用户成功访问过的 Web 页内容,可以对网络访问进行事后监督,为过滤机制的完善提供相应数据。文中分析了 Web 页的访问流程,基于 HTTP 代理服务器,在应用层实现了对 Web 页的关键字过滤和基于语义的内容过滤,并通过将客户机成功访问过的 Web 页存储在代理服务器硬盘上,实现了内容重现。试验表明,语义过滤能较好地甄别文本的不同观点,准确度较单纯关键字过滤有明显提高。

关键词:内容过滤;代理;内容重现;语义
中图分类号:TP301.2 **文献标识码**:A **文章编号**:1673-629X(2007)04-0120-05

Semantic Filtering and Content Recurrence for Web Page Accessing Based on Proxy

ZHOU Wen-gang, MA Zhan-xin
(Department of Computer Science, Zhoukou Normal University, Zhoukou 466001, China)

Abstract: It is important to construct a healthy and secure network circumstance through necessary and effective content filtering to Web page accessing. To recur the content of Web pages which have been accessed by users successfully can do postmortem supervising for Internet surfing, and to provide the corresponding data for consummating the filtering system. This paper analyzed the flow of the Web page accessing, the keyword-based filtering and the semantic-based content filtering for Web page on the application layer were implemented based on HTTP proxy server. Besides, the content recur function is implemented by saving the content of Web pages which are accessed successfully by client PC, to the HDD of the proxy server. The results of experiments proved that semantic filtering can discriminate the different standpoints, its precision is improved in evidence compared with content filtering only by key word.

Key words: content filtering; proxy; content recurrence; semantic

1 Web 页内容过滤的现状

Internet 上的 Web 页往往融会了文字、图片等多种媒体的资源,这使得信息的表达方式具有多样性。要实现对某一信息的完全过滤,理想的算法应能对该信息的每一种表达形式做出较为准确的匹配判断,这一要求目前还远没有达到,实际应用的 Web 访问过滤技术还仅停留在 URL 过滤、文本内容过滤和图像过滤上[1]。从实际效果看,URL 过滤速度快但容易以偏盖全;文本内容过滤常使用基于关键字的匹配判断,这种技术只能实现结构对应层次上的判断,基于语义的文本内容过滤技术目前也并不成熟[2];至于图像过滤,因为同一主体在不同图像中的表现可能有较大差异,很难单纯通过非智能计算来提取共同点,从而使图像过滤技术仅在某些特殊领域有所突破。

过滤技术的现状说明过滤准确度不可能完美地符合人们的要求,因此实际过滤操作中人们是在过滤速度和准确度之间寻求一种平衡。文中的重点在于介绍如何在 Web 页访问过程中实施必要的过滤。

2 HTTP 协议分析与 Web 页访问处理流程

使用 HTTP 进行网络访问的流程如图 1 所示。

①Web 服务器在指定端口侦听来自客户机的连接请求。

②用户向浏览器提交目标 Web 页的 URL。

③HTTP 客户机与 URL 指定的服务器建立 TCP 连接。

④HTTP 客户机向 HTTP 服务器发送请求,以期

收稿日期:2006-06-22
基金项目:河南省教育厅自然科学研究计划(2006520022);周口师范学院青年基金项目(ZKNUQN200615)
作者简介:周文刚(1972-),男,河南沈丘人,硕士,研究方向为网络安全;马占欣,副教授,研究方向为数据挖掘。

获取指定的 Web 页。

⑤HTTP 服务器根据请求的成功或失败发出相应的应答信息。如果请求成功,应答信息会包含所请求 Web 页的部分或全部内容。

⑥服务器发送应答信息后关闭连接。

⑦Web 浏览器处理客户机接收的应答信息并反馈请求结果给用户。

用户借助浏览器来访问 Web 页,一个 Web 浏览器软件至少包括 HTML 解释器和用来检索 HTML Web 页的 HTTP 客户程序(客户机)(见图 1),图 1 中的 HTTP 客户机和 HTTP 服务器对用户是透明的。用户向浏览器提交 URL 之后,浏览器和相应的 Web 服务器之间开始一次③~⑦的流程,一般把③~⑥的过程称作一次会话(session)。如果步骤⑤返回的 Web 页中有图片、声音等其他超级链接,HTTP 客户机会自动对每一个链接开始一次②~⑦的流程以下载它。因此,要完成一个完整的 Web 页显示,一般需要建立若干次②~⑦的流程[3]。

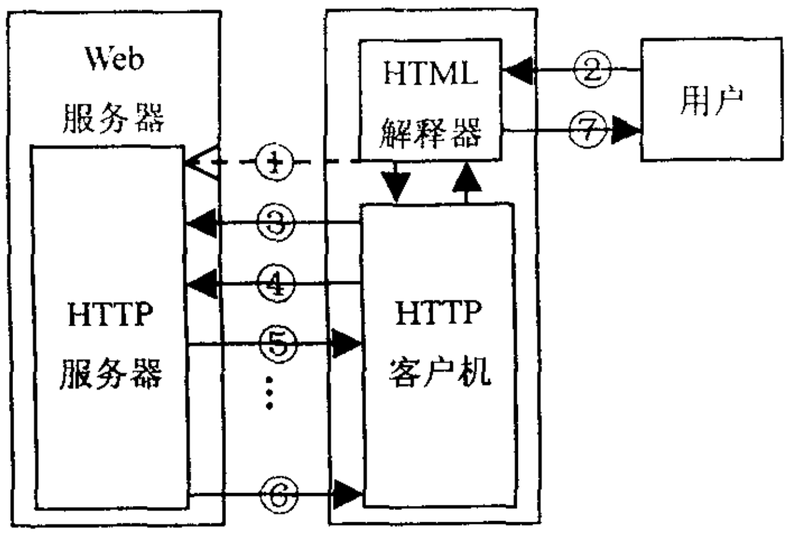

图 1　使用 FTTP 进行网络访问的流程

3　具有过滤功能的代理服务器

3.1　代理服务器的作用与工作流程

分析访问 Web 页的处理流程不难发现,在步骤④请求成功之后,步骤⑤会返回包含 Web 页源代码(部分或全部)的应答信息,之后交由浏览器进行 HTML 解释处理。如果在步骤⑤之后先进行过滤,则显然可以让浏览器的 HTML 解释器处理过滤后的网页信息。但如图 1 所示,用来接收应答信息的 HTTP 客户机与 HTML 解释器之间有着较高的耦合度,在两者之间插入过滤功能模块的可行性与效率还有待论证。

工作在应用层的代理服务器可以解决这一问题。作为“中介”,代理服务器分别与客户机和目标服务器建立 TCP 连接,按图 1 所示流程实现转发客户机的请求和目标服务器的应答,图 2 显示了一种简化的、带有过滤功能的代理服务器工作流程。

代理服务器做以下工作:

①在指定端口侦听并接收客户机对目标服务器的访问请求。

②将进行过 URL 或 IP 地址过滤的访问请求转发到目标服务器,被过滤的请求不予转发。

③接收目标服务器反馈的应答信息,送过滤模块进行文本过滤、图像过滤等操作。

④将过滤后的应答信息转发给客户机。

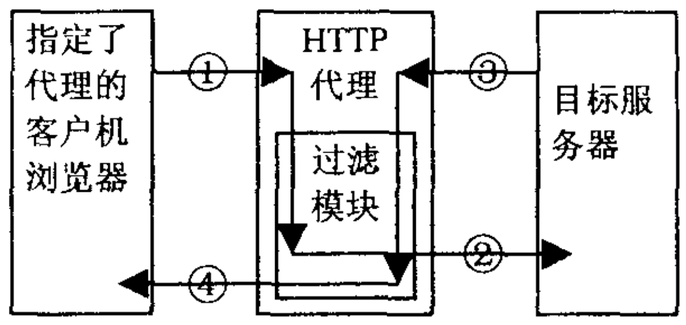

图 2　简化的、带有过滤功能的代理服务工作流程

显然,在图 2 所示的网络结构中,处于中枢地位的代理服务器是整个网络的性能瓶颈。面对众多客户机和服务器并发传送过来的访问请求和应答信息,单纯的转发已经耗费了代理服务器的大量资源,因此,附加的过滤模块如果没有高效(速)的过滤算法,将会因 Web 页转发的过度延迟而失去实用价值。

3.2　过滤模块设计

由图 2,要想让被过滤的 Web 页内容不被客户机的浏览器显示,切断①→②,③→④这两个环节中的任何一个都可以实现。显然,先切断①→②环节,则可以不再进行代理服务器与目标服务器之间基于 HTTP 的网络访问流程(图 1),从而提高代理效率。因此,过滤模块应该采用分级过滤的原则,即首先在①→②环节进行基于 URL 或 IP 地址的预过滤,然后在③→④环节进行 Web 页内容过滤。

3.2.1　基于域名或 IP 地址列表的客户端请求过滤

根据 HTTP1.1 协议,客户机发出的请求信息中包含了目标 Web 页的域名。为此,过滤模块中基于域名或 IP 地址列表的客户端请求过滤算法可描述如下:

①获取客户机发送的请求信息。

②从请求信息中的请求行或 Referer 报头域中取出域名,并根据域名求出对应的 IP 地址。

③检查域名或 IP 地址是否在过滤列表中,若在,代理服务器回答一个拒绝访问的应答信息给客户机,中止此次会话;否则,转发该请求到目标服务器。

过滤列表中的域名和 IP 地址应是动态的。

3.2.2　Web 页文本内容过滤

对 Web 页内容的过滤主要是指对组成该 Web 页的文本、脚本、图片等元素进行过滤。Web 页的源代码是一个纯字符集,因此对于 Web 页文本内容过滤和脚本过滤都是基于字符串比较的基本技术。

Web 页源码包含了大量的 HTML 标记符,事先将非 HTML 标记符字符集中存储到一个字符串变量 m_

WebChar 中,当所要过滤的关键字多于一个时可减少关键字匹配运算时的比较次数。

1)简单的关键字匹配过滤算法。

简单的关键字匹配过滤算法就是判断 m_ WebChar 中是否包含指定的过滤关键字。代理服务器提供动态设置过滤关键字的接口,生成过滤关键字的集合 m_Keyword 数组,通过统计 m_ Keyword 数组中的每个元素在 m_ WebChar 中出现的次数来决定目标 Web 页是否应过滤。这种过滤算法只能机械地实现结构对应层次的匹配判断,考虑不到关键字所处的上下文语境。而且,为了回避过滤,Web 页的设计者有多种方法可以改变关键字的表达方式,比如在关键字的字符间夹杂其它字符,也可以把关键字用图形来显示,诸如此类的变化方式过滤模块是无法穷尽的。

简单的关键字匹配过滤算法误判率较高,其过滤结果并不可信,但是其过滤速度快,故在一些领域还有着广泛使用。

2)基于语义的过滤算法[2]。

对 Web 页的文字内容应用基于语义的信息过滤可以实现较高的查准率[4~6]。其算法简述如下:

①定义能描述语义框架的数据结构,用来表述语句中的行为主体、受体、行为,以及行为发生的时间、地点,不妨设语义框架的槽的集合为 $\{s_1, s_2, \cdots, s_n, \text{weight}\}$。

② 接收欲过滤的语句字符串做模板,对其进行分词,根据分词结果(词的词性和在句中的位置) 填充语义框架中各槽(s_i) 的值,生成语义框架 F_0,按公式(1)

$$\text{weight}(F) = \frac{\sum_{i=1}^{n} w(s_i) f_{\text{dist}}(s_i)}{\sum_{i=1}^{n} w(s_i)} \tag{1}$$

计算 F_0 的匹配权重并存入 weight 槽中。其中 $f_{\text{dist}}(s_i)$ 由公式(2) 定义,它描述槽 s_i(特征项) 在语句中与行为动词之间的距离关系。$w(s_i)$ 由公式(3) 定义,它描述 s_i 在语句中的关键程度。

$$f_{\text{dist}}(s) = \begin{cases} 1 & \text{在同一个句子中} \\ 0.5 & \text{在同一个取样窗口中} \\ 0.25 & \text{在同一个段落中} \\ 0.1 & \text{其他情况} \end{cases} \tag{2}$$

$$W(s_i) = \begin{cases} 2, & s_i = \text{行为主体、受体} \\ 3, & s_i = \text{施加的行为,中心动词} \\ 1, & s_i = \text{行为发生时间、地点} \end{cases} \tag{3}$$

③ 将描述行为主体、受体、行为的词添加到关键字列表 m_ Keyword 数组中。

④ 检查目标 Web 页中是否包含 m_ Keyword 数组中的词,如果不包含,直接放行,此次过滤判断结束;否则,将包含关键字的语句赋值给字符串变量 Sentence。

⑤ 对 Sentence 中字符串进行分词,并生成语义框架 F_1,按公式(1) 计算其匹配权重。

⑥ 按公式(4) 计算框架 F_0 与 F_1 之间的相似度 $\text{Sim}(F_0, F_1)$。

$$\text{sim}(F_0, F_1) = \left(\sum_{i=1}^{n} \text{similar}(F_0, s_i, F_1, s_i)\right) * F_0\,\text{weight} * F_1\text{weight} \tag{4}$$

其中,

$$\text{similar}(S_1, S_2) = \begin{cases} 1, & S_1 \text{ 与 } S_2 \text{ 完全相同} \\ 0.75, & S_1 \text{ 与 } S_2 \text{ 仅词性相同} \\ 0, & S_1 \text{ 与 } S_2 \text{ 完全不相同} \end{cases}$$

如果 $\text{Sim}(F_0, F_1)$ 的值大于给定的阈值,该 Web 页予以过滤,发拒绝访问信息至客户机,终止此次会话;否则转 ④,对 Web 页的剩余内容继续判断。

语义过滤算法中分词是关键工作,分词效果的好坏决定着语义过滤能否成功。文献[2] 中给出了基于词库的最大匹配分词算法,试验表明,分词的正确率和速度满足过滤的需要。

4 基于代理的网络访问内容重现算法

需要过滤的信息和 Web 页设计者应对过滤的手段是不断发展变化的,因此过滤设置应该是个动态的过程。实际应用中经常有某信息应被过滤但却被放行的现象发生,故应该建立相应的反馈机制以完善过滤系统。

内容重现就是将客户机通过代理服务器访问过的 Web 页面及该网页所包含的所有类型的文件保存下来。实施内容重现有以下好处:

(1)存储常访问 Web 页可以提高代理速度[7]。

(2)可帮助系统管理员查找过滤判断失误的原因。

(3)可以为非法访问留存证据。

(4)为挖掘客户访问行为和偏好提供数据储备。

实现内容重现的算法如下:

①将目标 Web 页的 URL 按下述规则转换成在代理服务器上存放的路径名,并生成相应的空文件。

对于不含 Web 页文件名的 URL,如 http://www.sohu.com/,反转成相对路径名 www.sohu.com\xx.html,其中 xx 取当前的系统时间来替代,以免在对同一网站多次访问时造成文件存储冲突。

对于包含文件名的 URL,直接取“http://”之后的字符串反转成相对路径,并增加系统时间信息以避免存储时文件名冲突。

②将 URL 与转换后的文件名的对应关系存入访问日志。

③代理服务器在转发目标服务器的应答信息给客户机时,将去掉 HTTP 的应答头的应答信息(网页内容)存储到步骤①所生成的文件中。

5　在内容重现的基础上改进代理技术

为了节省存储空间,当对同一 URL 再次访问时,判断响应信息中的 Last - Modified 值是否晚于访问日志中保存的该 URL 的 Last - Modified 值,如果相同,则仅在访问日志中增加一条指向原存储位置的信息,对 Web 页内容不再重复存储。同时,代理服务器关闭与目标服务器建立的此次会话,并将存储本机上的对应文件内容附加 HTTP 响应头后发给客户机,这样可以提高代理服务器的工作效率。

6　用 VC + + 实现 Web 访问内容过滤与重现

6.1　网络环境设置

双穴主机一台做服务器,其网卡 A 连接外网,网卡 B 连接内网,为网卡 B 绑定内网专用 IP。为服务器的浏览器(以 IE 为例)设置代理,地址为网卡 B 绑定的 IP,端口号任意。客户机若干,网卡 IP 设为与服务器网卡 B 在同一网段的专用 IP,网关设为服务器网卡 B 的 IP,为浏览器设置代理,地址和端口同代理服务器浏览器的代理设置。

6.2　代理服务器软件的实现

由图 2,代理服务器软件的关键功能包括侦听、与客户机和目标服务器进行信息交换和过滤等。下面给出用 Visual C + + 设计代理服务器软件的主要步骤:

(1)定义一个存储会话信息的结构体数据 Session,其成员变量用来描述一次会话所产生的客户机端和服务器端的套接字句柄、IP、端口号、以及所传递的请求信息或响应信息的内容等数据。

代理服务器在进行信息转发时,由于网速慢等原因,有时需要将会话信息保存一段时间。为方便会话信息的快速读写与共享,可以创建一个内存映射文件用来按 Session 结构存储所有的会话信息。

(2)定义三个派生自 CAsyncSocket 类的类 CListen、CClient 和 CServer。CListen 类对象负责在指定端口(必须是代理服务器在 IE 中设置的代理端口)侦听来自客户机的连接请求,CClient 类对象主要负责接收来自客户机的请求信息,CServer 类对象主要负责建立与目标服务器的连接并转发客户机的请求信息,也负责接收来自目标服务器的响应信息。

重载 CListen::OnAccept(),实现当侦听到客户机的连接信息时,创建一个 CClient 类对象,建立与客户机的连接,并将接收请求信息的工作交由 CClient:OnReceive()来处理,然后返回侦听状态。

重载 CClient:OnReceive(),实现接收客户机发出的请求信息,并对请求信息进行域名和 IP 过滤判断,如果在过滤之列,发拒绝信息到客户机,并终止此次会话;否则,创建一个 CServer 类对象,通过它转发请求信息到目标服务器,同时将此次会话的有关数据存储到步骤(1)中建立的内存映射文件中。处理流程如下:

①创建一对套接字对象(CClient 的 m_pSocket 和 CServer 的 m_pServerSocket);

②接收客户端的请求头信息,存入 SRequire 中;

③从 SRequire 取出 URL 和端口,并将 URL 转换成 IP 地址;

④进行 URL 或 IP 地址过滤判断,如果符合过滤条件,发拒绝信息给客户端,终止此次会话,不向目标服务器转发此请求;

⑤取得客户端 IP 地址和端口;

⑥利用 m_pServerSocket 创建一个自动选择端口的转发套接字,向目标服务器的指定端口发出连接请求;

⑦创建一个会话,将当前套接字有关信息存入 m_Session 中去。参数包括客户端套和服务器端的套接字句柄、IP、端口号,URL 和请求头等;

重载 CServer::OnReceive(),实现从目标服务器中接收响应信息,根据当前套接字的句柄在内存映射文件查找对应的记录,如果没找到,终止此次会话;否则,取出其中的客户端套接字句柄 m_pClient,对响应信息进行关键字和基于语义的内容过滤判断,如果需过滤,就通过 m_pClient 向客户机发拒绝信息,否则,就通过 m_pClient 向客户机转发此响应信息。

CServer::OnReceive()按下述流程工作:

①创建一对套接字对象(CClient 和 CServer);

②按当前套接字句柄查找会话记录中有无该句柄;

③如果存在该句柄但对应的客户端套接字为 NULL:就关闭该服务套接字,并删除会话记录中相应数据;

④如果存在该句柄并且对应的客户端套接字不为 NULL,则:

创建互斥量,获得互斥量的存取权;

读取来自目标服务器端的应答信息(可能包含网页内容)存入 RBuf 中;

对 RBuf 中数据进行关键字、脚本、URL、语义等过滤判断,如果符合过滤条件,将该会话记录的 bAction 属性值置为 FR_DENY;

⑤如果某会话记录的 bAction 属性值为 FR_DE-

NY:

取出该记录中的客户端套接字,通过它向客户机发拒绝信息;

取出服务器套接字,关闭它;

否则:

如果该会话记录中的客户端套接字不为 NULL:

通过该客户端套接字转发应答信息到客户机;

否则:

关闭该客户端套接字;

⑥释放互斥量;

⑦存储成功转发的应答信息到代理服务器硬盘,以便实现内容重现。

对响应信息进行关键字和基于语义的内容过滤判断,如果需过滤,就通过 m_pClient 向客户机发拒绝信息,否则,就通过 m_pClient 向客户机转发此响应信息。

7 性能分析

在 Windows XP SP2 环境下,用 Visual C++ 6.0 实现了上述算法。在 10 台客户机和一台代理服务器(均为 P4/2G/256M)搭建的子网环境中试验。取 100 篇 Web 页(文章),其中设置对某问题的赞成、反对、中立观点各 25 篇。应用简单的关键字过滤,查全率达到 98%,应用基于语义的过滤,根据设置的过滤语句模板,可以较好地甄别出 3 种观点,查准率达 93.5%。在内容重现的基础上对代理技术改进之后,对常用站点的访问速度明显加快。

参考文献:

[1] 马文斌,王 庆. Web 内容过滤实现方法的研究[J]. 计算机工程,2004(12):588-589.

[2] 周文刚,王景中. 基于语义的信息过滤算法的设计和实现[J]. 周口师范学院学报,2006,23(2):99-100.

[3] 张明武,陈启祥,楚惟善,等. HTTP 代理服务系统的实现与分析[J]. 计算机工程,2001,27(3):145-147.

[4] Policová G, Návrat P. Semantic Similarity in Content-Based Filtering[C]//In: Advances in Databases and Information Systems: 6th East European Conference, ADBIS 2002. Bratislava, Slovakia: [s. n.], 2002.

[5] Jorge J, Flores G. Semantic Filtering of Textual Requirements Descriptions[C]//In: 9th International Conference on Applications of Natural Language to Information Systems, NLDB 2004. Salford, UK: [s. n.], 2004.

[6] YU Fei, SHEN Yue, AN Ji-yao, et al. Information Audit Based on Image Content Filtering[J]. Wuhan University Journal of Natural Sciences, 2006, 11(1): 234-238.

[7] 李家欣,倪 亮,王 乘. 具备高速缓存的 HTTP 代理防火墙的设计与实现[J]. 计算机工程,2003,29(3):85-86.

(上接第 119 页)

的过滤策略,增加邮件过滤的针对性。

4) 便于功能扩充,系统管理者可自主增加、停用或修改过滤引擎,进行自定义的功能扩展。

5) 配置方便,各个模块分工明确,系统管理者可根据邮件用户需求定制使用或是不使用各个过滤模块。

在实现上,方案中存在的关键问题有:

(1) 对邮件内容进行规则匹配、贝叶斯运算、检测病毒都会占用大量的系统资源。

(2) 由于邮件过滤模块的层次较多,各过滤方式之间过滤内容上有重复、叠加的情况,导致过滤效率降低。

对问题(1),可以将病毒过滤模块和邮件过滤模块分别部署在不同主机上,将各模块采取分布式结构加以解决。对问题(2),可以不断优化策略管理来减少各个过滤方式在过滤内容上的重复过滤。

3 结束语

在计算机信息安全领域,垃圾邮件的泛滥成为严峻的现实,带毒邮件的检测过滤更是成为垃圾邮件过滤的新挑战,不同用户的不同需求也应得到个性体现和满足。基于综合方法的邮件过滤方案的提出,在一定程度上反映了反垃圾邮件的安全和个性化需求,也是对现有过滤方法的有益补充。

参考文献:

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[EB/OL]. 2006-01. http://www.cnnic.net.cn.

[2] RFC706. On the junk mail problem[S]. 1975.

[3] 潘文峰. 基于内容的垃圾邮件过滤研究[D]. 北京:中国科学院,2004.

[4] Graham. A Plan for Spam[EB/OL]. 2002-08. http://www.paulgraham.com/spam.html.

[5] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Evaluation of Naive Bayesian Anti-Spam Filtering[C]//Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000). Barcelona, Spain: [s. n.], 2000.

[6] Symantec. Understanding Polymorphic Viruses[EB/OL]. 2005-08. http://www.symantec.com.

[7] 刘尊全. 计算机病毒防范与信息对抗技术[M]. 北京:清华大学出版社,1991.