(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102882703 B
(45)授权公告日 2015.08.19

(21)申请号 201210320169.6

(22)申请日 2012.08.31

(73)专利权人 赛尔网络有限公司
地址 100084 北京市海淀区中关村东路1号院清华科技园8号楼B座赛尔大厦

(72)发明人 何旭 李威 黄友俊 李星 吴建平

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 宋焰琴

(51)Int.Cl.
H04L 12/24(2006.01)
H04L 12/26(2006.01)
H04L 29/08(2006.01)

(56)对比文件

CN 101453424 A,2009.06.10,第6页第1段至第7页第26行,第8页第14至第18行,第9页13行至16行.

CN 102055620 A,2011.05.11,说明书第[0111]段至第[0116]段,第[0119]段至第[0123]段,附图1-3.

CN 101453424 A,2009.06.10,第6页第1段至第7页第26行,第8页第14至第18行,第9页13行至16行.

WO 02/099688 A1,2002.12.12,全文.

WO 2006/077454 A1,2006.07.27,全文.

CN 102394885 A,2011.03.28,全文.

US 2008/0059508 A1,2008.04.06,全文.

CN 101872347 A,2010.10.27,全文.

审查员 杜宇坤

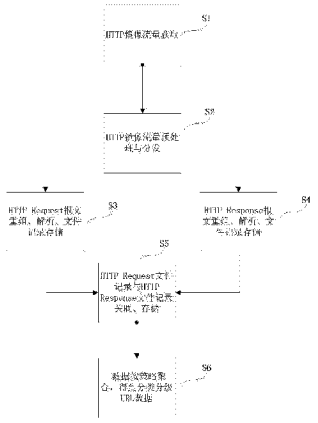
权利要求书3页 说明书8页 附图4页

(54)发明名称

一种基于HTTP分析的URL自动分类分级的系统及方法

(57)摘要

本发明公开了一种基于HTTP分析的URL自动分类分级的系统和方法,该系统包括用于分发HTTP请求/响应报文镜像数据流的HTTP请求/响应报文预处理器、用于对HTTP请求报文镜像数据流进行重组的HTTP请求报文解析服务器、用于对HTTP响应报文镜像数据流进行重组的HTTP响应报文解析服务器、交换机、用于存储报文信息的HTTP存储关联服务器和用于对URL进行自动分类分级的HTTP分级分类服务器。该方法包括以下步骤:预处理HTTP报文镜像数据流得到HTTP请求/响应报文镜像数据流,并对其进行分发;对请求/响应报文镜像数据流进行重组,将从重组数据中提取到的报文信息存储起来;对存储的信息进行关联;根据关联信息对URL分类分级。本发明能够实现对URL的高效的分类分级。



CN 102882703 B

1. 一种基于 HTTP 分析的 URL 自动分类分级系统，其特征在于，该系统包括 HTTP 请求 / 响应报文预处理器、HTTP 请求报文解析服务器、HTTP 响应报文解析服务器、交换机、HTTP 存储关联服务器和 HTTP 分级分类服务器，其中：

所述 HTTP 请求 / 响应报文预处理器用于接收 HTTP 报文镜像数据流并对其进行预处理得到 HTTP 请求报文镜像数据流和 HTTP 响应报文镜像数据流，并将所述 HTTP 请求报文镜像数据流发给所述 HTTP 请求报文解析服务器，将 HTTP 响应报文镜像数据流发给所述 HTTP 响应报文解析服务器，其中，所述预处理为：根据 TCP 层的端口号，将目的端口号为 80、8080、443 的 HTTP 报文镜像数据流标识为 HTTP 请求报文镜像数据流，将源端口号为 80、8080、443 的 HTTP 报文镜像数据流标识为 HTTP 响应报文镜像数据流；

所述 HTTP 请求报文解析服务器与所述 HTTP 请求 / 响应报文预处理器连接，用于对所接收的 HTTP 请求报文镜像数据流进行重组处理，从重组后的数据流中提取报文信息，并将所述报文信息以 HTTP 请求文件记录的形式存储到所述 HTTP 存储关联服务器上；

所述 HTTP 响应报文解析服务器与所述 HTTP 请求 / 响应报文预处理器连接，用于对所接收的 HTTP 响应报文镜像数据流进行重组处理，从重组后的数据流中提取报文信息，并将所述报文信息以 HTTP 响应文件记录的形式存储到所述 HTTP 存储关联服务器上；

所述交换机与 HTTP 请求 / 响应报文预处理器、HTTP 请求报文解析服务器、HTTP 响应报文解析服务器、HTTP 存储关联服务器、HTTP 分类分级服务器相连，用于对 HTTP 请求 / 响应报文预处理器、HTTP 请求报文解析服务器、HTTP 响应报文解析服务器及运行其上的程序进行远程配置管理，同时也作为 HTTP 请求报文解析服务器、HTTP 响应报文解析服务器、HTTP 存储关联服务器、HTTP 分类分级服务器之间的数据传输通道；

所述 HTTP 存储关联服务器与所述交换机相连，用于根据所述报文信息中的五元组信息，将所述 HTTP 请求文件记录与所述 HTTP 响应文件记录进行关联，并将关联上的数据以记录格式按行存入存储文件；

所述 HTTP 分类分级服务器与所述交换机相连，用于根据用户定制的策略，通过所述交换机访问所述 HTTP 关联存储服务器中的记录信息，对 URL 进行分类分级。

2. 根据权利要求 1 所述的系统，其特征在于，所述五元组信息包括：客户端 IP 地址 client_ip、服务器 IP 地址 server_ip、客户端端口号 client_port、服务器端口号 server_port 和关联序列号 rel_seqno。

3. 一种基于 HTTP 分析的 URL 自动分类分级方法，其特征在于，该方法包括以下步骤：

步骤 S1，在骨干网的路由器上获取 HTTP 报文镜像数据流；

步骤 S2，HTTP Request/Response 报文预处理器对所述 HTTP 报文镜像数据流进行预处理得到 HTTP 请求报文镜像数据流和 HTTP 响应报文镜像数据流，并将所述 HTTP 请求报文镜像数据流发送给 HTTP 请求报文解析服务器，将 HTTP 响应报文镜像数据流发送给 HTTP 响应报文解析服务器，其中，所述预处理为：根据 TCP 层的端口号，将目的端口号为 80、8080、443 的 HTTP 报文镜像数据流标识为 HTTP 请求报文镜像数据流，将源端口号为 80、8080、443 的 HTTP 报文镜像数据流标识为 HTTP 响应报文镜像数据流；

步骤 S3，所述 HTTP 请求报文解析服务器对所接收的 HTTP 请求报文镜像数据流进行重组处理，并将从重组后的数据流中提取到的报文信息以 HTTP 请求文件记录的形式存储到所述 HTTP 存储关联服务器上；

步骤 S4，所述 HTTP 响应报文解析服务器对所接收的 HTTP 响应报文镜像数据流进行重组处理，并将从重组后的数据流中提取到的报文信息以 HTTP 响应文件记录的形式存储到所述 HTTP 存储关联服务器上；

步骤 S5，HTTP 存储关联服务器根据由 client_ip、server_ip、client_port、server_port、rel_seqno 组成的五元组信息，将所述 HTTP 请求文件记录与所述 HTTP 响应文件记录进行关联，并将关联上的数据以记录格式按行存入存储文件；

步骤 S6，HTTP 分类分级服务器根据用户定制的策略，访问所述 HTTP 关联存储服务器中的记录信息，将 URL 进行分类分级。

4. 根据权利要求 3 所述的方法，其特征在于，所述步骤 S3 中提取的报文信息包括：客户端 IP 地址 client_ip、服务器 IP 地址 server_ip、客户端端口号 client_port、服务器端口号 server_port、关联序列号 rel_seqno、主机 Host、请求 URL 请求 -URL；所述步骤 S4 中提取的报文信息包括：客户端 IP 地址 client_ip、服务器 IP 地址 server_ip、客户端端口号 client_port、服务器端口号 server_port、关联序列号 rel_seqno、内容类型 Content-Type、内容编码 Content-Encoding、内容语言 Content-Language、内容长度 Content-Length。

5. 根据权利要求 3 所述的方法，其特征在于，所述步骤 S3 中，所述 HTTP 请求报文解析服务器对所述 HTTP 请求报文镜像数据流进行重组处理进一步包括以下步骤：

步骤 S300，获取所述 HTTP 请求报文镜像数据流的当前 TCP 分片；

步骤 S310，解析所述 HTTP 请求报文的 IP/TCP 报首，从中获取客户端 IP 地址 client_ip，服务器 IP 地址 server_ip，客户端端口号 client_port，服务器端口号 server_port，ack 序列号 ack_seqno，根据获取的这些信息计算出下一报文序列号 next_seqno 和关联序列号 rel_seqno；根据 HTTP 请求报文的报首特征，判断该 TCP 分片是否为 HTTP 请求报文报首的第一个 TCP 分片，若是，则继续判断 HTTP 请求报文的报首是否完整，如果完整则进入步骤 S320，否则进入步骤 S330；如果该 TCP 分片不是 HTTP 请求报文报首的第一个 TCP 分片，则以 client_ip，server_ip，client_port，server_port，ack_seqno 为索引查找预先存储的 HTTP 流表，判断所述 HTTP 流表中是否有与所述索引匹配的流表表项，如果是则进入步骤 S340，否则结束本次流程，回到步骤 S300 进入下一次流程；

步骤 S320，解析所述 HTTP 请求报文的报首，从中提取出 HTTP 版本号 HTTP Version，请求 URI Request-URI，主机 Host；将所述步骤 S310 得到的 client_ip，server_ip，client_port，server_port，rel_seqno，Host，Request-URI 以 HTTP 请求文件记录的形式写入 HTTP 关联存储服务器，并删除所述匹配的 HTTP 流表表项，结束本次流程，回到步骤 S300 进入下一次流程；

步骤 S330，以 client_ip，server_ip，client_port，server_port，ack_seqno 为索引，新建一个 HTTP 流表表项，结束本次流程，回到步骤 S300 进入下一次流程；

步骤 S340，对与所述索引匹配的 HTTP 流表表项进行 TCP 重组，然后再判断所述 HTTP 请求报文的报首是否完整，若完整则返回步骤 S320，否则结束本次流程，回到步骤 S300 进入下一次流程。

6. 根据权利要求 3 所述的方法，其特征在于，所述步骤 S4 中，所述 HTTP 响应报文解析服务器对所述 HTTP 响应报文镜像流进行重组处理进一步包括以下步骤：

步骤 S400，获取所述 HTTP 响应报文镜像数据流的当前 TCP 分片，进入步骤 S410 ；

步骤 S410，解析所述 HTTP 响应报文镜像数据流的 IP/TCP 报首，从中获取客户端 IP 地址 client_ip，服务器 IP 地址 server_ip，客户端端口号 client_port，服务器端端口号 server_port，ack 序列号 ack_seqno，根据获取的这些信息计算下一报文序列号 next_seqno 和关联序列号 rel_seqno ；根据 HTTP 响应报文的报首特征，判断该 TCP 分片是否为 HTTP 响应报文报首的第一个 TCP 分片，若是，则继续判断 HTTP 响应报文的报首是否完整，如果完整则进入步骤 S420，否则进入步骤 S430 ；如果该 TCP 分片不是 HTTP 响应报文报首的第一个 TCP 分片，则以 client_ip，server_ip，client_port，server_port，ack_seqno 为索引查找预先存储的 HTTP 流表，判断所述 HTTP 流表中是否有与所述索引匹配的流表表项，如果是则进入步骤 S440，否则结束本次流程，回到步骤 S400 进入下一次流程 ；

步骤 S420，解析所述 HTTP 响应报文的报首，从中提取出状态码 Status-Code，内容类型 Content-Type，内容长度 Content-Length，内容编码 Content-Encoding，内容语言 Content-Language，并将所述步骤 S410 得到的 client_ip，server_ip，client_port，server_port，rel_seqno，Status-Code，Content-Type，Content-Length，Content-Encoding，Content-Language 以 HTTP 响应文件记录的形式写入 HTTP 关联存储服务器，并删除所述匹配的 HTTP 流表表项，结束本次流程，回到步骤 S400 进入下一次流程 ；

步骤 S430，以 client_ip，server_ip，client_port，server_port，ack_seqno 为索引，新建一个 HTTP 流表表项，结束本次流程，回到步骤 S400 进入下一次流程 ；

步骤 S440，对与所述索引匹配的 HTTP 流表表项进行 TCP 重组，然后再判断所述 HTTP 响应报文的报首是否完整，若完整则返回步骤 S420，否则结束本次流程，回到步骤 S400 进入下一次流程。

7. 根据权利要求 3 所述的方法，其特征在于，所述记录格式包含如下字段 ：client_ip，server_ip，Host，URL，Content-Type，Content-Length，Content-Encoding，Content-Language。

8. 根据权利要求 4 所述的方法，其特征在于，所述步骤 S6 进一步包括以下步骤 ：

步骤 S61，从 HTTP 请求文件记录中提取出属性信息 URL、Host ；

步骤 S62，从 HTTP 响应文件记录中提取出属性信息 Content-Type、Content-Encoding、Content-Language、Content-Length ；

步骤 S63，根据属性信息 Content-Type、Content-Encoding、Content-Language、Content-Length、Host 对 URL 进行分级和分类。

一种基于 HTTP 分析的 URL 自动分类分级的系统及方法

技术领域

[0001] 本发明涉及网络行为监控与网络行为管理技术领域，更具体地，涉及一种基于 HTTP 分析的 URL 自动分类分级的系统及方法。

背景技术

[0002] 据互联网追踪机构 Netcraft 在 2011 年 10 月 9 日的统计报告显示：全球网站总量约 5 亿，其中真正处于活动状态的仅为 1.5 亿。根据数据分析：目前全球网站总数庞大，“垃圾网站”超过 50％，并且处于增长态势，互联网环境有待于清理及净化。在中国社会科学院 2011 年发布的《新媒体蓝皮书》中显示：2010 年中国互联网站总数达 191 万，网页数量 600 亿。

[0003] 众所周知，互联网上的网页和所有其他资源都是通过 URL 标识的，而网络资源访问的一半以上是通过 HTTP 协议承载的。面对如此众多的 URL，仅凭人工标识达到分类分级的目的显然是不现实的。

[0004] 现有常用的 HTTP 报文的 TCP 分片重组算法是：将 HTTP 协议的 TCP 分片按照 src_ip、dst_ip、src_port、server_ip 四元组匹配，并且以 SYN 报文的 seqno 作为起始序号，以 FIN 报文的 seqno 作为结束序号，进行 TCP 流的跟踪和重组，得到一个完整的 TCP 流后，再对上层的 HTTP 协议进行解析。这样做的缺点是：(1)HTTP/1.1 标准中，一个 TCP 流中可以包含多次的 HTTP 的 Request 与 Response，对 HTTP 解析提取增加的判断复杂性。(2)HTTP Response 报文可能承载着音视频数据，造成 TCP 流的持续时间很长，增加了系统的时间与空间的开销。

发明内容

[0005] 为了解决上述现有技术存在的缺陷，本发明提出一种基于 HTTP 分析的 URL 自动分类分级的系统及方法。该方法可以独立对 URL 进行分类分级，也可以结合人工标识对 URL 进行分类分级，并且还可以作为预处理阶段的方法。

[0006] 根据本发明的一方面，提出一种基于 HTTP 分析的 URL 自动分类分级系统，其特征在于，该系统包括 HTTP 请求 / 响应报文预处理器、HTTP 请求报文解析服务器、HTTP 响应报文解析服务器、交换机、HTTP 存储关联服务器和 HTTP 分级分类服务器，其中：

[0007] 所述 HTTP 请求 / 响应报文预处理器用于接收 HTTP 报文镜像数据流并对其进行预处理得到 HTTP 请求报文镜像数据流和 HTTP 响应报文镜像数据流，并将所述 HTTP 请求报文镜像数据流发给所述 HTTP 请求报文解析服务器，将 HTTP 响应报文镜像数据流发给所述 HTTP 响应报文解析服务器；

[0008] 所述 HTTP 请求报文解析服务器与所述 HTTP 请求 / 响应报文预处理器连接，用于对所接收的 HTTP 请求报文镜像数据流进行重组处理，从重组后的数据流中提取报文信息，并将所述报文信息以 HTTP 请求文件记录的形式存储到所述 HTTP 存储关联服务器上；

[0009] 所述 HTTP 响应报文解析服务器与所述 HTTP 请求 / 响应报文预处理器连接，用于

对所接收的 HTTP 响应报文镜像数据流进行重组处理，从重组后的数据流中提取报文信息，并将所述报文信息以 HTTP 响应文件记录的形式存储到所述 HTTP 存储关联服务器上；

[0010] 所述交换机与 HTTP 请求 / 响应报文预处理器、HTTP 请求报文解析服务器、HTTP 响应报文解析服务器、HTTP 存储关联服务器、HTTP 分类分级服务器相连，用于对 HTTP 请求 / 响应报文预处理器、HTTP 请求报文解析服务器、HTTP 响应报文解析服务器及运行其上的程序进行远程配置管理，同时也作为 HTTP 请求报文解析服务器、HTTP 响应报文解析服务器、HTTP 存储关联服务器、HTTP 分类分级服务器之间的数据传输通道；

[0011] 所述 HTTP 存储关联服务器与所述交换机相连，用于根据所述报文信息中的五元组信息，将所述 HTTP 请求文件记录与所述 HTTP 响应文件记录进行关联，并将关联上的数据以记录格式按行存入存储文件；

[0012] 所述 HTTP 分类分级服务器与所述交换机相连，用于根据用户定制的策略，通过所述交换机访问所述 HTTP 关联存储服务器中的记录信息，对 URL 进行分类分级。

[0013] 根据本发明的另一方面，提出一种基于 HTTP 分析的 URL 自动分类分级方法，其特征在于，该方法包括以下步骤：

[0014] 步骤 S1，在骨干网的路由器上获取 HTTP 报文镜像数据流；

[0015] 步骤 S2，HTTP Request/Response 报文预处理器对所述 HTTP 报文镜像数据流进行预处理得到 HTTP 请求报文镜像数据流和 HTTP 响应报文镜像数据流，并将所述 HTTP 请求报文镜像数据流发送给 HTTP 请求报文解析服务器，将 HTTP 响应报文镜像数据流发送给 HTTP 响应报文解析服务器；

[0016] 步骤 S3，所述 HTTP 请求报文解析服务器对所接收的 HTTP 请求报文镜像数据流进行重组处理，并将从重组后的数据流中提取到的报文信息以 HTTP 请求文件记录的形式存储到所述 HTTP 存储关联服务器上；

[0017] 步骤 S4，所述 HTTP 响应报文解析服务器对所接收的 HTTP 响应报文镜像数据流进行重组处理，并将从重组后的数据流中提取到的报文信息以 HTTP 响应文件记录的形式存储到所述 HTTP 存储关联服务器上；

[0018] 步骤 S5，HTTP 存储关联服务器根据由 client_ip、server_ip、client_port、server_port、rel_seqno 组成的五元组信息，将所述 HTTP 请求文件记录与所述 HTTP 响应文件记录进行关联，并将关联上的数据以记录格式按行存入存储文件；

[0019] 步骤 S6，HTTP 分类分级服务器根据用户定制的策略，访问所述 HTTP 关联存储服务器中的记录信息，将 URL 进行分类分级。

[0020] 根据上述本发明的技术方案，本发明的有益效果为：(1) 只针对已识别的 HTTP 报文的第一个 TCP 分片建立流表表项，以 client_ip、server_ip、client_port、server_ip、ack_seqno 为索引，后续 TCP 分片根据索引查找对应的表项，并根据 seqno 进行排列重组；(2) 以 HTTP 报首和数据的分隔符为重组结束标志，对于 HTTP Response 报文，由于只关注报首的重组，从而对于持续时间长的 TCP 流，大大地节省了系统时间与空间的开销。

附图说明

[0021] 图 1 为本发明基于 HTTP 分析的 URL 自动分类分级系统结构图。

[0022] 图 2 为本发明基于 HTTP 分析的 URL 自动分类分级方法流程图。

[0023] 图 3 为本发明 HTTP Request 报文报首重组解析逻辑图。

[0024] 图 4 为本发明 HTTP Response 报文报首重组解析逻辑图。

具体实施方式

[0025] 为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明进一步详细说明。在描述过程中省略了对于本发明来说是不必要的细节和功能，以防止对本发明的理解造成混淆。

[0026] 图 1 为本发明基于 HTTP 分析的 URL 自动分类分级系统结构图，如图 1 所示，根据本发明的一方面，提出一种基于 HTTP 分析的 URL 自动分类分级系统，该系统包括 :HTTP 请求 / 响应 Request/Response 报文预处理器、HTTP 请求 Request 报文解析服务器、HTTP 响应 Response 报文解析服务器、交换机、HTTP 存储关联服务器和 HTTP 分级分类服务器，其中 :

[0027] 所述 HTTP Request/Response 报文预处理器用于接收所述 HTTP 报文镜像数据流并对其进行预处理得到 HTTP Request 报文镜像数据流和 HTTPResponse 报文镜像数据流，并将所述 HTTP Request 报文镜像数据流发给 HTTP Request 报文解析服务器，将 HTTP Response 报文镜像数据流发给 HTTP Response 报文解析服务器。具体地，所述 HTTP Request/Response 报文预处理器具有 4 个网络接口，网络接口 1 用于接收 HTTP 报文镜像数据流，并对其进行预处理 :根据 TCP 层的端口号，将目的端口号为 80、8080、443 的 HTTP 报文镜像数据流标识为 HTTP Request 报文镜像数据流，将源端口号为 80、8080、443 的 HTTP 报文镜像数据流标识为 HTTP Response 报文镜像数据流 ;网络接口 2 用于将所述 HTTP Request 报文镜像数据流发送给所述 HTTP Request 报文解析服务器 ;网络接口 3 用于将所述 HTTP Response 报文镜像数据流发送给所述 HTTP Response 报文解析服务器 ;网络接口 4 与所述交换机连接，用于对 HTTP Request/Response 报文预处理器及运行其上的程序进行配置管理。

[0028] 所述 HTTP Request 报文解析服务器与所述 HTTP Request/Response 报文预处理器连接，用于对所接收的 HTTP Request 报文镜像数据流进行重组处理，从重组后的数据流中提取报文信息，并将所述报文信息以 HTTP 请求文件记录的形式存储到所述 HTTP 存储关联服务器上 ;所述 HTTP Request 报文解析服务器具有 2 个网络接口，网络接口 1 与 HTTP Request/Response 报文预处理器的网络接口 2 相连，用于接收所述 HTTP Request 报文镜像数据流，并对其进行重组处理，然后将重组处理后的所述 HTTP Request 报文进行解析，提取出客户端 IP 地址 (client_ip)、服务器 IP 地址 (server_ip)、客户端端口号 (client_port)、服务器端口号 (server_port)、关联序列号 (rel_seqno)、主机 (Host)、请求 URL(Request-URL) 等报文信息，并将提取出的上述报文信息以 HTTP Request 文件记录的形式通过与所述交换机相连的网络接口 2 存储到所述 HTTP 存储关联服务器上，另外，还可通过所述网络接口 2 对 HTTP Request 报文解析服务器及运行其上的程序进行远程配置管理 ;

[0029] 所述 HTTP Response 报文解析服务器与所述 HTTP Request/Response 报文预处理器连接，用于对所接收的 HTTP Response 报文镜像数据流进行重组处理，从重组后的数据流中提取报文信息，并将所述报文信息以 HTTP 响应文件记录的形式存储到所述 HTTP 存储关联服务器上 ;所述 HTTP Response 报文解析服务器具有 2 个网络接口，网络接口 1 与

HTTP Request/Response 报文预处理器的网络接口 3 相连，用于接收所述 HTTPResponse 报文镜像数据流，并对其进行重组处理，然后将重组处理后的所述 HTTP Response 报文进行解析，提取出客户端 IP 地址 (client_ip)、服务器 IP 地址 (server_ip)、客户端端口号 (client_port)、服务器端口号 (server_port)、关联序列号 (rel_seqno)、内容类型 (Content-Type)、内容编码 (Content-Encoding)、内容语言 (Content-Language)、内容长度 (Content-Length) 等报文信息，并将提取出的上述报文信息以 HTTP Response 文件记录的形式通过与所述交换机相连的网络接口 2 存储到所述 HTTP 存储关联服务器上，另外，还可通过所述网络接口 2 对 HTTP Response 报文解析服务器及运行其上的程序进行远程配置管理；

[0030] 所述交换机进一步为通讯千兆交换机，所述通讯千兆交换机与 HTTPRequest/Response 报文预处理器、HTTP Request 报文解析服务器、HTTPResponse 报文解析服务器、HTTP 存储关联服务器、HTTP 分类分级服务器的配置管理网络接口相连，用于使系统维护人员对 HTTP Request/Response 报文预处理器、HTTP Request 报文解析服务器、HTTP Response 报文解析服务器及运行其上的程序进行远程配置管理，另外也作为 HTTP Request 报文解析服务器、HTTP Response 报文解析服务器、HTTP 存储关联服务器、HTTP 分类分级服务器之间的数据传输通道。

[0031] 所述 HTTP 存储关联服务器与所述交换机相连，用于根据所述报文信息中的 client_ip、server_ip、client_port、server_port、rel_seqno 五元组信息，将所述 HTTP Request 文件记录与所述 HTTP Response 文件记录进行关联，并将关联上的数据以记录格式按行存入存储文件。所述 HTTP 存储关联服务器具有 1 个网络接口，所述 HTTP 存储关联服务器通过该网络接口与所述交换机连接，用于与 HTTP Request 报文解析服务器、HTTPResponse 报文解析服务器、HTTP 分类分级服务器之间进行数据传输，并可通过该网络接口对 HTTP 存储关联服务器及运行其上的程序进行远程配置管理。

[0032] 所述 HTTP 分类分级服务器与所述交换机相连，用于根据用户定制的策略，通过所述交换机访问所述 HTTP 关联存储服务器中的记录信息，对 URL 进行分类分级，HTTP 分类分级服务器具有 1 个网络接口，所述 HTTP 分类分级服务器通过该网络接口与所述交换机连接，用于与 HTTP 存储关联服务器之间进行数据传输，并可通过该网络接口对 HTTP 分类分级服务器及运行其上的程序进行远程配置管理。

[0033] 图 2 为本发明基于 HTTP 分析的 URL 自动分类分级方法流程图，如图 2 所示，根据本发明的另一方面，还提出一种基于 HTTP 分析的 URL 自动分类分级方法，该方法包括以下步骤：

[0034] 步骤 S1，在骨干网的路由器上获取 HTTP 报文镜像数据流；

[0035] 步骤 S2，HTTP Request/Response 报文预处理器对所述 HTTP 报文镜像数据流进行预处理得到 HTTP Request 报文镜像数据流和 HTTPResponse 报文镜像数据流，并将所述 HTTP Request 报文镜像数据流发送给 HTTP Request 报文解析服务器，将 HTTP Response 报文镜像数据流发送给 HTTP Response 报文解析服务器；

[0036] 所述预处理进一步为：根据 TCP 层的端口号，将目的端口号为 80、8080、443 的 HTTP 报文镜像数据流标识为 HTTP Request 报文镜像数据流，将源端口号为 80、8080、443 的 HTTP 报文镜像数据流标识为 HTTP Response 报文镜像数据流。

[0037] 步骤 S3，所述 HTTP Request 报文解析服务器对所接收的 HTTP Request 报文镜像数据流进行重组处理，然后将重组处理后的所述 HTTP Request 报文进行解析，提取出 client_ip、server_ip、client_port、server_port、rel_seqno、Host、Request-URI 等报文信息，并将提取出的上述报文信息以 HTTP Request 文件记录的形式存储到 HTTP 存储关联服务器中；

[0038] 该步骤中，所述 HTTP Request 报文解析服务器对所述 HTTP Request 报文镜像数据流进行重组处理进一步包括以下步骤（如图 3 所示）：

[0039] 步骤 S300，获取所述 HTTP Request 报文镜像数据流的当前 TCP 分片，进入步骤 S310；

[0040] 步骤 S310，解析所述 HTTP Request 报文镜像数据流的 IP/TCP 报首，从中获取客户端 IP 地址 client_ip，服务器 IP 地址 server_ip，客户端端口号 client_port，服务器端口号 server_port，ack 序列号 ack_seqno 等信息，根据获取的这些信息计算出下一报文序列号 next_seqno 和关联序列号 rel_seqno；根据 HTTP Request 报文的报首特征，判断该 TCP 分片是否为 HTTP Request 报文报首的第一个 TCP 分片，若是，则继续判断 HTTPRequest 报文的报首是否完整，如果完整则进入步骤 S320，否则进入步骤 S330；如果该 TCP 分片不是 HTTP Request 报文报首的第一个 TCP 分片，则以 client_ip，server_ip，client_port，server_port，ack_seqno 为索引查找预先存储的 HTTP 流表，判断所述 HTTP 流表中是否有与所述索引匹配的流表表项，如果是则进入步骤 S340，否则结束本次流程，回到步骤 S300 进入下一次流程；

[0041] 对于 HTTP 报首的开始与结束的判断方法，可参考 RFC2068 标准文档。

[0042] 其中，采用如下公式根据获取的 client_ip，server_ip，client_port，server_port，ack_seqno 等信息计算出 next_seqno，rel_seqno：

[0043] next_seqno = seqno+payload_length，rel_seqno = next_seqno，

[0044] 其中，seqno 为 TCP 分片报首中的序列号，payload_length 为 TCP 有效载荷长度。

[0045] 所述根据 HTTP Request 报文的报首特征，判断该 TCP 分片是否为 HTTP Request 报文报首的第一个 TCP 分片进一步为：以" \r\n" 为行结束符，从 TCP 分片的数据段中提取首行数据，将其与 HTTP Requset 报文的请求行的正则表达式" GET.*HTTP./." 进行匹配，若匹配成功，则判断该 TCP 分片是 HTTP Request 报文报首的第一个 TCP 分片；若否，则不是。

[0046] 步骤 S320，解析所述 HTTP Request 报文的报首，从中提取出 HTTP 版本号 HTTP Version，请求 URI Request-URI，主机 Host 等信息；将所述步骤 S310 得到的 client_ip，server_ip，client_port，server_port，rel_seqno 以及主机 Host、Request-URI 等信息以 HTTP Request 文件记录的形式写入 HTTP 关联存储服务器，并删除所述步骤 S310 中匹配的 HTTP 流表表项，结束本次流程，回到步骤 S300 进入下一次流程；

[0047] 步骤 S330，以 client_ip，server_ip，client_port，server_port，ack_seqno 为索引，新建一个 HTTP 流表表项，结束本次流程，回到步骤 S300 进入下一次流程；

[0048] 每个 HTTP 流表表项包含两个数据结构：一个链表 List<TcpSegment> 和一个搜索二叉树 Tree<seqno，TcpSegment>。所述链表用于存放已重组好的 TCP 分片；所述搜索二叉树用于存放未重组的 TCP 分片，并且以 TCP 分片的 seqno 作为搜索二叉树的键值 key。

[0049] 在步骤 S330 中，新建一个 HTTP 流表表项时，链表 List<TcpSegment> 和搜索二叉树 Tree<seqno，TcpSegmet> 均为空，将该 TCP 分片放入链表 List<TcpSegment> 的首部，并回到步骤 S300。

[0050] 步骤 S340，对与所述索引匹配的 HTTP 流表表项进行 TCP 重组，重组后再判断所述 HTTP Request 报文的报首是否完整，若完整则返回步骤 S320，否则结束本次流程，回到步骤 S300 进入下一次流程。

[0051] 步骤 S340 中，对与索引匹配的 HTTP 流表表项进行 TCP 重组进一步为：

[0052] 如果找到一个与该 TCP 分片的索引匹配的 HTTP 流表表项，则判断该 TCP 分片的 seqno 是否等于 List<TcpSegment> 链表尾部的 TCP 分片的下一报文序列号 next_seqno：如果二者相等，那么就将该 TCP 分片加入 List<TcpSegment> 链表的尾部，并且遍历 Tree<seqno，TcpSegmet> 搜索二叉树，对搜索二叉树中的每一个 TCP 分片重复上述比较，直至遍历完整个 Tree<seqno，TcpSegmet> 搜索二叉树或者在相应的 TCP 分片的数据段中匹配到字符串″ \r\n\r\n″，如果是匹配到字符串″ \r\n\r\n″，那么就将链表中存储的 TCP 分片的数据段重组为完整的 HTTP Request 报文，并删除相应的 HTTP 流表表项，返回步骤 S320；如果直至遍历完整个 Tree<seqno，TcpSegmet> 搜索二叉树也没有匹配到字符串″ \r\n\r\n″，则直接回到步骤 S300；如果二者不等，则直接回到步骤 S300。

[0053] 所述从重组处理后的数据流中提取出 client_ip、server_ip、client_port、server_port、rel_seqno、Host、Request-URI 等信息进一步为：

[0054] 从重组处理后的数据流中提取出源 IP 地址 src_ip，并进一步从 src_ip 中得到 client_ip，提取出目的 IP 地址 dst_ip，并进一步从 dst_ip 中得到 server_ip，提取出源端口号 src_port，并进一步从 src_port 中得到 client_port，提取出目的端口号 dst_port，并进一步从 dst_port 中得到 server_port，将最后一个 TCP 分片的序列号 seqno 加上 TCP 有效载荷长度 payload_length 得到 rel_seqno。

[0055] 步骤 S4，所述 HTTP Response 报文解析服务器对所接收的 HTTP Response 报文镜像数据流进行重组处理，然后将重组处理后的所述 HTTP Response 报文进行解析，并提取出 client_ip、server_ip、client_port、server_port、rel_seqno、Content-Type、Content-Encoding、Content-Language、Content-Length 等信息，将提取出的信息以 HTTP Response 文件记录的形式存储到所述 HTTP 存储关联服务器中；

[0056] 该步骤中，所述 HTTP Response 报文解析服务器对所述 HTTP Response 报文镜像流进行重组处理进一步包括以下步骤（如图 4 所示）：

[0057] 步骤 S400，获取所述 HTTP Response 报文镜像数据流的当前 TCP 分片，进入步骤 S410；

[0058] 步骤 S410，解析所述 HTTP Response 报文镜像数据流的 IP/TCP 报首，从中获取客户端 IP 地址 client_ip，服务器 IP 地址 server_ip，客户端端口号 client_port，服务器端口号 server_port，ack 序列号 ack_seqno 等信息，根据获取的这些信息计算下一报文序列号 next_seqno 和关联序列号 rel_seqno；根据 HTTP Response 报文的报首特征，判断该 TCP 分片是否为 HTTP Response 报文报首的第一个 TCP 分片，若是，则继续判断 HTTP Response 报文的报首是否完整，如果完整则进入步骤 S420，否则进入步骤 S430；如果该 TCP 分片不是 HTTP Response 报文报首的第一个 TCP 分片，则以 client_ip，server_ip，client_port，

server_port,ack_seqno为索引查找预先存储的HTTP流表,判断所述HTTP流表中是否有与所述索引匹配的流表表项,如果是则进入步骤S440,否则结束本次流程,回到步骤S400进入下一次流程;

[0059] 其中,计算下一报文序列号next_seqno和关联序列号rel_seqno的公式、对于第一个TCP分片的判断方法均与步骤S310类似,在此不做赘述。

[0060] 步骤S420,解析所述HTTP Response报文的报首,从中提取出状态码Status-Code,内容类型Content-Type,内容长度Content-Length,内容编码Content-Encoding,内容语言Content-Language等信息,并将所述步骤S410得到的client_ip,server_ip,client_port,server_port,rel_seqno以及状态码Status-Code,内容类型Content-Type,内容长度Content-Length,内容编码Content-Encoding,内容语言Content-Language等信息以HTTPResponse文件记录的形式写入HTTP关联存储服务器,并删除所述步骤S410中匹配的HTTP流表表项,结束本次流程,回到步骤S400进入下一次流程;

[0061] 步骤S430,以client_ip,server_ip,client_port,server_port,ack_seqno为索引,新建一个HTTP流表表项,结束本次流程,回到步骤S400进入下一次流程;

[0062] 在步骤S430中,新建一个HTTP流表表项时,链表List<TcpSegment>和搜索二叉树Tree<seqno,TcpSegmet>均为空,将该TCP分片放入链表List<TcpSegment>的首部,并回到步骤S400。

[0063] 步骤S440,对与所述索引匹配的HTTP流表表项进行TCP重组,重组后再判断所述HTTP Response报文的报首是否完整,若完整则返回步骤S420,否则结束本次流程,并回到步骤S400进入下一次流程。

[0064] 步骤S440中,对与索引匹配的HTTP流表表项进行TCP重组进一步为:

[0065] 如果找到一个与该TCP分片的索引匹配的HTTP流表表项,则判断该TCP分片的seqno是否等于List<TcpSegment>链表尾部的TCP分片的下一报文序列号next_seqno:如果二者相等,那么就将该TCP分片加入List<TcpSegment>链表的尾部,并且遍历Tree<seqno,TcpSegmet>搜索二叉树,对搜索二叉树中的每一个TCP分片重复上述比较,直至遍历完整个Tree<seqno,TcpSegmet>搜索二叉树或者在相应的TCP分片的数据段中匹配到字符串"\r\n\r\n",如果是匹配到字符串"\r\n\r\n",那么就将链表中存储的TCP分片的数据段重组为完整的HTTPResponse报文,并删除相应的HTTP流表表项,返回步骤S420;如果直至遍历完整个Tree<seqno,TcpSegmet>搜索二叉树也没有匹配到字符串"\r\n\r\n",则直接回到步骤S400;如果二者不等,则直接回到步骤S400。

[0066] 所述从重组处理后的数据流中提取出client_ip、server_ip、client_port、server_port、rel_seqno、Content-Type、Content-Encoding、Content-Language、Content-Length等信息进一步为:

[0067] 从重组处理后的数据流中提取出目的IP地址dst_ip,并进一步从dst_ip中得到client_ip,提取出源IP地址src_ip,并进一步从src_ip中得到server_ip,提取出目的端口号dst_port,并进一步从dst_port中得到client_port,提取出源端口号src_port,并进一步从src_port中得到server_port,提取出确认序列号ack_seqno,并进一步从ack_seqno中得到rel_seqno。

[0068] 步骤S5，HTTP存储关联服务器根据由client_ip、server_ip、client_port、server_port、rel_seqno组成的五元组信息，将所述HTTP Request文件记录与所述HTTP Response文件记录进行关联，并将关联上的数据以记录格式按行存入存储文件；

[0069] 该步骤中，在将HTTP Request文件记录与HTTP Response文件记录进行关联时，首先过滤掉状态码Status-Code不等于200的数据，然后再将Host，Request-URI拼接成完整的URL，并将关联上的数据以记录格式按行存入存储文件，记录格式包含如下字段：client_ip，server_ip，Host，URL，Content-Type，Content-Length，Content-Encoding，Content-Language。

[0070] 步骤S6，HTTP分类分级服务器根据用户定制的策略，访问所述HTTP关联存储服务器中的记录信息，将URL进行分类分级。

[0071] 匹配成功的HTTP Request文件记录和HTTP Response文件记录对应一次完整的HTTP交互。从HTTP Request文件记录中可以提取出URL、Host等属性信息，从HTTP Response文件记录中可以提取出Content-Type、Content-Encoding、Content-Language、Content-Length等属性信息，通过根据Content-Type、Content-Encoding、Content-Language、Content-Length、Host等属性信息可以对URL进行分级、分类。而HTTP报文报首的解析与文件记录关联、属性提取、根据属性分级分类都可以在人为制定策略后由计算机程序完成，从而达到自动化的目的。

[0072] 根据对一段时间内HTTP报文报首的解析与关联结果的数据分析，可以得到不同纬度的URL分级与分类，并对URL打上相应的标签：比如，可以根据Host将URL按照所在网站分类；根据Content-Type将URL按照内容类型分类；根据Content-Encoding将URL按压缩类型分类；根据Content-Language将URL按照语言类型分类；根据server_ip将URL按照所处网段分类；根据Content-Length将URL按照内容大小分级；根据单位时间内URL被访问次数，将URL按照热点程度分级；或结合以上的一种或多种进行多维度的分类分级。

[0073] 比如URL按照网站分类的标签可能有sina.com、google.com、bupt.edu.cn等，URL按Content-Type分类的标签可能有text、video、audio、image等，URL按Content-Languague分类的标签可能有English、Chinese、Japanese等，URL按照访问次数分级的标签可能有每天被访问10次以下、10-100次、100-1000次、1000-10000次、10000次以上，当用户希望搜索热点为每天被访问10000次以上的语言为中文的热点帖子时，就可以通过定制Content-Type标签为text、Content-Languague标签为Chinese、访问次数在10000次以上，HTTP分类分级服务器就会根据这些标签条件从存储文件中搜索出符合的URL及相关信息的记录。

[0074] 以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

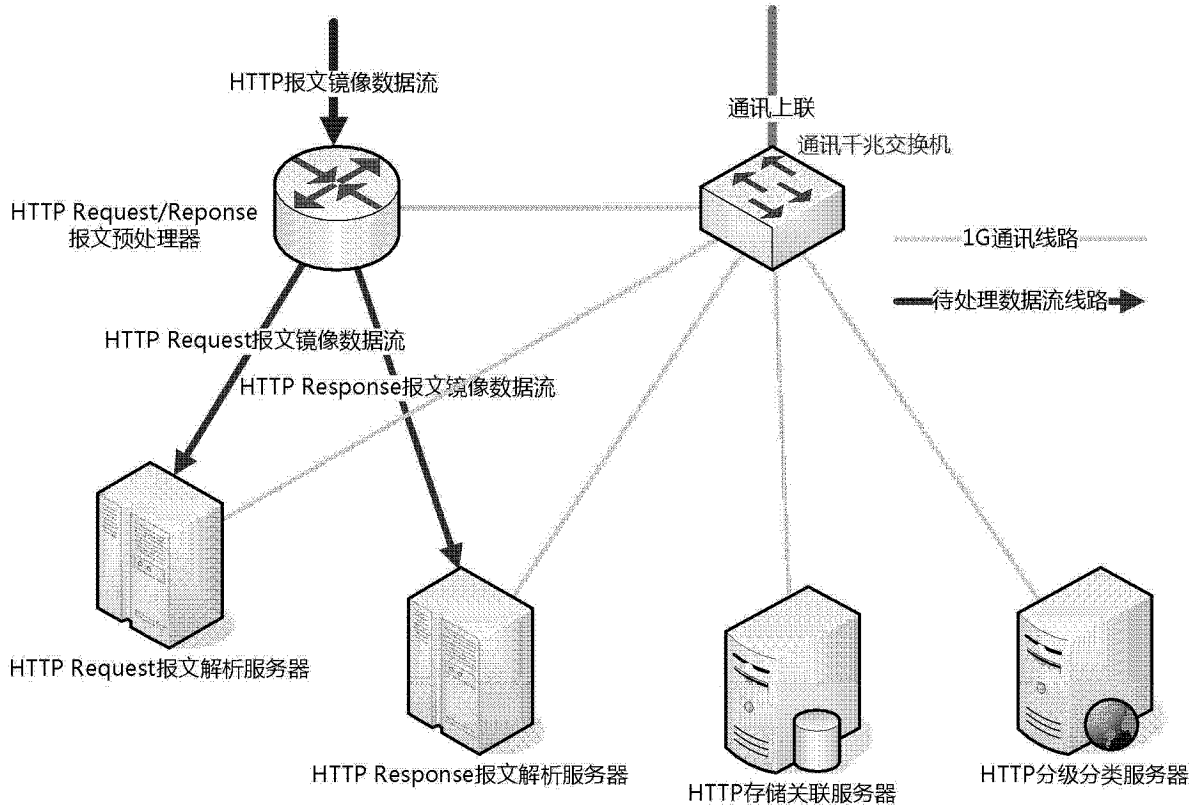

HTTP报文镜像数据流
通讯上联
通讯千兆交换机
HTTP Request/Reponse
报文预处理器
1G通讯线路
待处理数据流线路
HTTP Request报文镜像数据流
HTTP Response报文镜像数据流
HTTP Request报文解析服务器
HTTP Response报文解析服务器
HTTP存储关联服务器
HTTP分级分类服务器

图 1

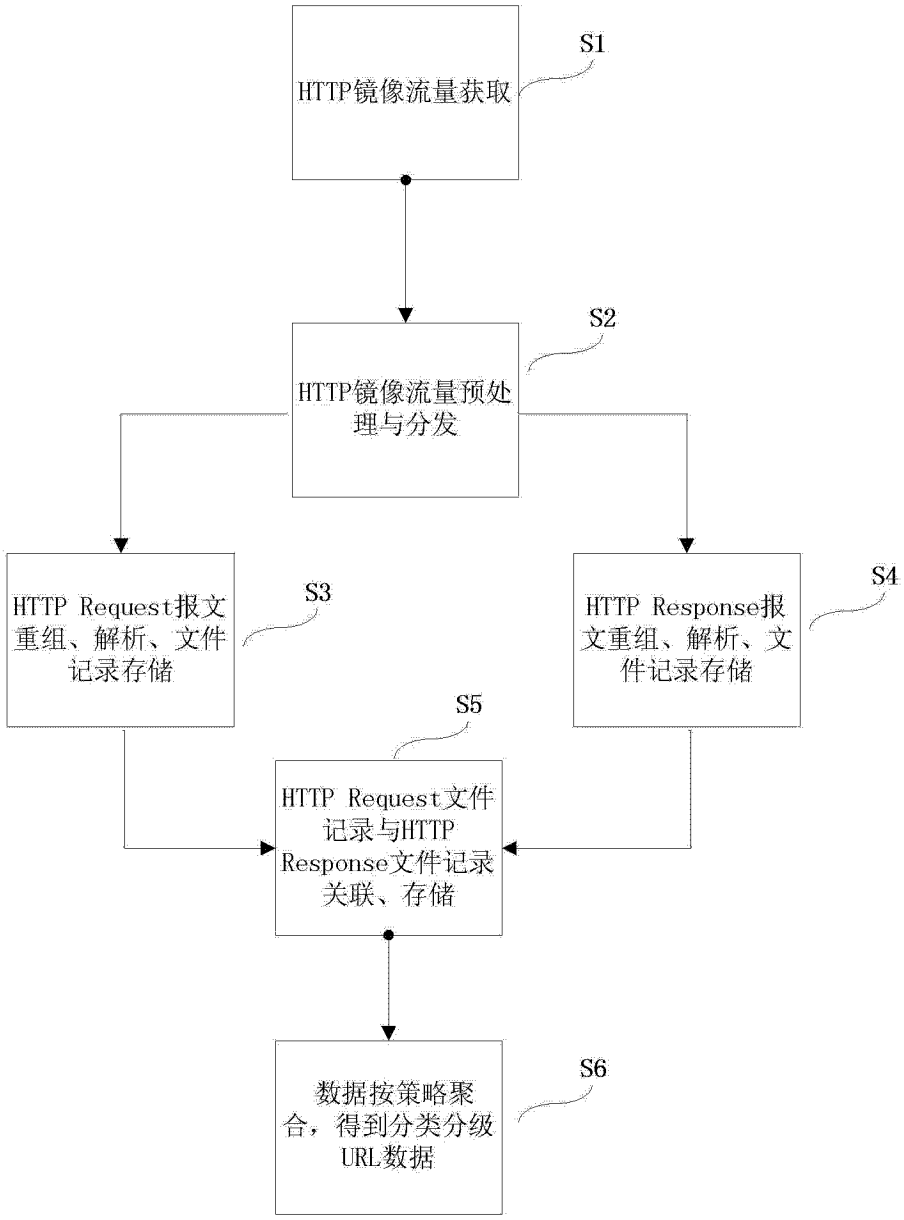

HTTP镜像流量获取
S1
HTTP镜像流量预处理与分发
S2
HTTP Request报文重组、解析、文件记录存储
S3
HTTP Response报文重组、解析、文件记录存储
S4
HTTP Request文件记录与HTTP Response文件记录关联、存储
S5
数据按策略聚合，得到分类分级URL数据
S6

图 2

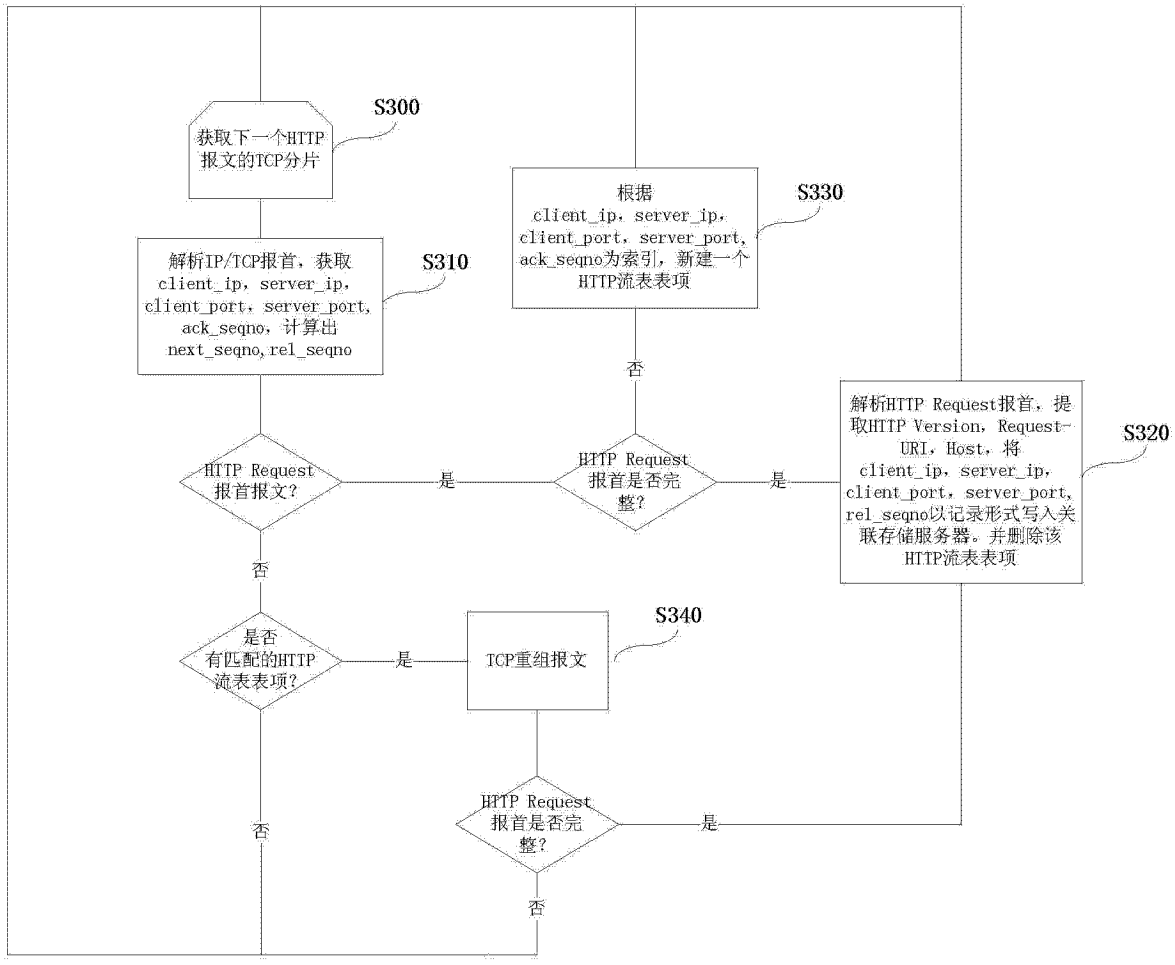

S300
获取下一个HTTP报文的TCP分片
S310
解析IP/TCP报首，获取client_ip，server_ip，client_port，server_port，ack_seqno，计算出next_seqno，rel_seqno
S330
根据client_ip，server_ip，client_port，server_port，ack_seqno为索引，新建一个HTTP流表表项
S320
解析HTTP Request报首，提取HTTP Version，Request-URI，Host，将client_ip，server_ip，client_port，server_port，rel_seqno以记录形式写入关联存储服务器。并删除该HTTP流表表项
S340
TCP重组报文
HTTP Request报首报文?
是
否
HTTP Request报首是否完整?
是
否
是否有匹配的HTTP流表表项?
是
否
HTTP Request报首是否完整?
是
否

图 3

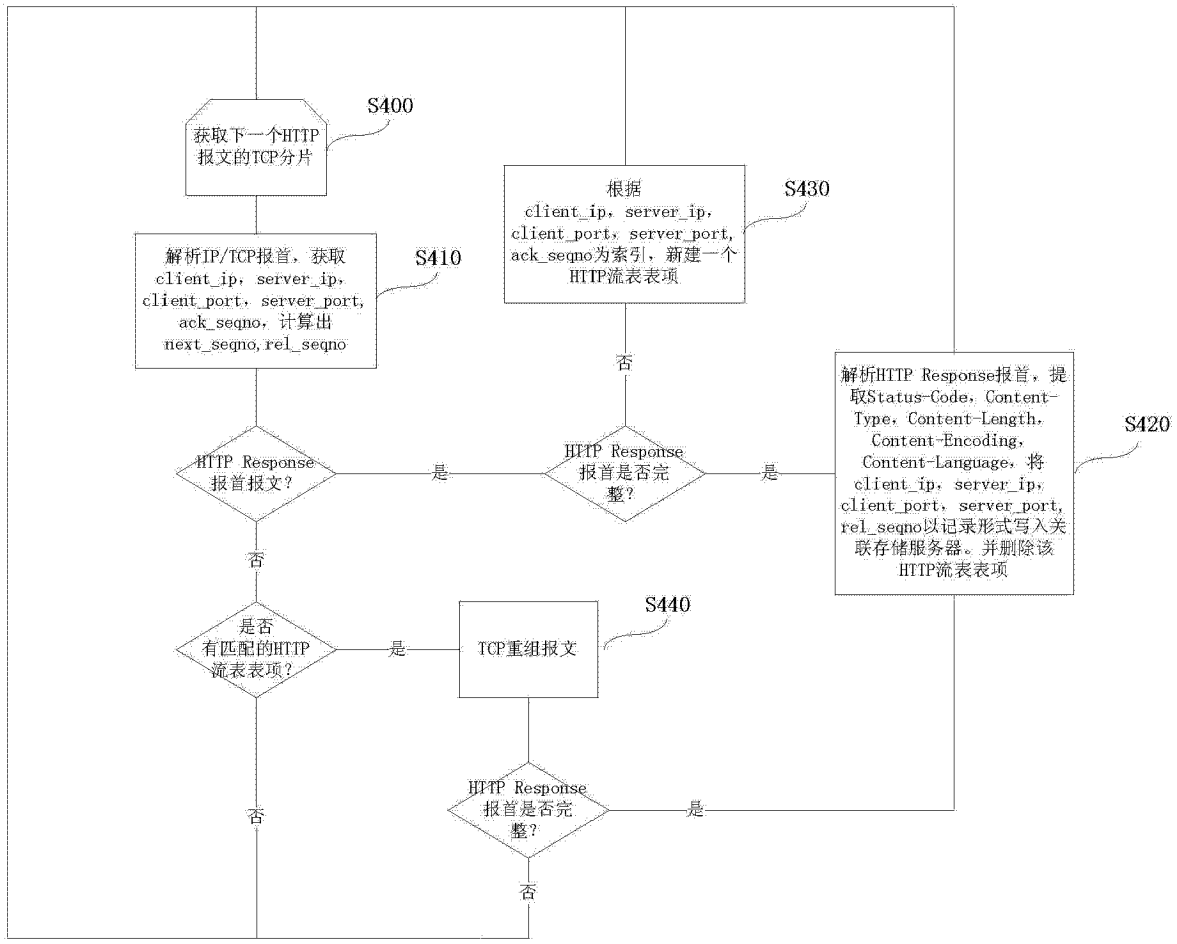

S400
获取下一个HTTP报文的TCP分片
S410
解析IP/TCP报首，获取client_ip，server_ip，client_port，server_port，ack_seqno，计算出next_seqno,rel_seqno
HTTP Response报首报文?
是
否
S430
根据client_ip，server_ip，client_port，server_port，ack_seqno为索引，新建一个HTTP流表表项
否
HTTP Response报首是否完整?
是
S420
解析HTTP Response报首，提取Status-Code，Content-Type，Content-Length，Content-Encoding，Content-Language，将client_ip，server_ip，client_port，server_port，rel_seqno以记录形式写入关联存储服务器。并删除该HTTP流表表项
是否有匹配的HTTP流表表项?
是
S440
TCP重组报文
否
HTTP Response报首是否完整?
是
否

图 4