(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2009 年2 月5 日 (05.02.2009)

PCT

(10) 国际公布号
WO 2009/015603 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2008/071811

(22) 国际申请日: 2008 年7 月30 日 (30.07.2008)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
200710075449.4
2007 年7 月30 日 (30.07.2007) CN

(71) 申请人 (对除美国外的所有指定国): 华为技术有限公司(HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为基地总部办公楼, Guangdong 518129 (CN)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): 王浩(WANG, Hao) [CN/CN]; 中国广东省深圳市龙岗区坂田华为基地总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 北京中博世达专利商标代理有限公司(BEIJING ZBSD PATENT & TRADEMARK AGENT LTD.); 中国北京市海淀区大柳树路17号富海大厦B座501室, Beijing 100081 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,

[见续页]

(54) **Title:** REGULAR EXPRESSION COMPILING SYSTEM, MATCHING SYSTEM, COMPILING METHOD AND MATCHING METHOD

(54) **发明名称:** 正则表达式编译、匹配系统及编译、匹配方法

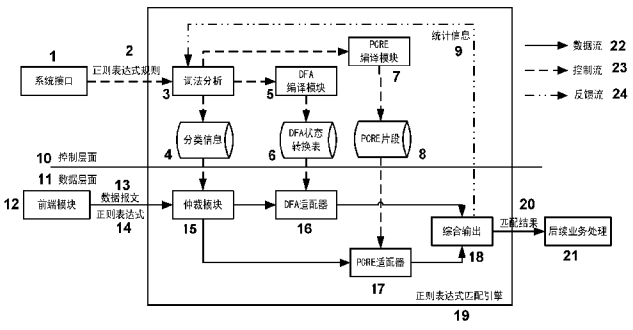

图2/Fig.2

1 SYSTEM INTERFACE
2 REGULAR EXPRESSION RULE
3 MORPHEME ANALYSIS
4 CLASSIFICATION INFORMATION
5 DFA COMPILATION MODULE
6 DFA STATUS CONVERSION TABLE
7 PCRE COMPILATION MODULE
8 PCRE SEGMENT
9 STATISTIC INFORMATION
10 CONTROL LAYER
11 DATA LAYER
12 FRONT MODULE
13 DATA PACKET
14 REGULAR EXPRESSION
15 JUDGMENT MODULE
16 DFA ADAPTER
17 PCRE ADAPTER
18 SYNTHESIS OUTPUT
19 REGULAR EXPRESSION MATCHING ENGINE
20 MATCHING RESULT
21 FOLLOWING SERVICE PROCESS
22 DATA FLOW
23 CONTROL FLOW
24 FEEDBACK FLOW

(57) Abstract: A regular expression compiling system, matching system, compiling method and matching method are disclosed. The regular expression compiling system provided by the embodiments of the present invention includes a morpheme analysis module and at least two kinds of compilation modules. The morpheme analysis module analyses the morphological characteristic of the regular expression, sends the regular expression to a corresponding compilation module for processing according to the previously set morphological rule and the morphological characteristic of the regular expression, and said corresponding compilation module receives the regular expression and compiles the regular expression to data structure with particular format.

[见续页]

WO 2009/015603 A1

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国（除另有指明，要求每一种可提供的地区保护）：ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告。

(57) 摘要：

本发明公开了一种正则表达式编译、匹配系统及编译、匹配方法。本发明实施例提供的的正则表达式编译系统包括词法分析模块和至少两种编译模块，所述词法分析模块分析正则表达式的词法特点，根据预设词法规则和所述正则表达式的词法特点将正则表达式送往相应的编译模块处理，所述相应的编译模块接收正则表达式，将所述正则表达式编译成特定形式的数据结构。

正则表达式编译、匹配系统及编译、匹配方法

技术领域

本发明属于通讯领域，尤其涉及一种正则表达式编译、匹配系统及编译、匹配方法。

背景技术

随着全 IP 网络（ALL-IP）、固定移动融合（Fixed and Mobile Convergence，简称 FMC）、多重播放（Triple-play）等概念的提出，传统的 IP 网络正在向融合数据、语音和视频于一体的多业务统一承载网转变。但是 IP 网络固有的数据传送方式和本质上开放的特征无法很好地满足电信级业务的需要，在网络安全性、可管理性和关键业务的 QoS 和 QoE 保证等方面都有待改进。为了对一些关键业务进行精确的业务识别和控制，除了按照传统的方法分析报文头中的五元组等字段外，还需要对报文的负荷部分进行检测，比如一些 P2P 流量使用非知名的端口，这样仅靠分析五元组就无法识别出此类流量。深度报文检测（Deep Packet Inspection，DPI）技术作为一种灵活有效的业务识别技术应运而生，并广泛应用在防火墙、IDS/IPS、业务控制网关等设备上，实现比如应用层负载均衡、基于特征的安全过滤等业务。和传统的报文检测方式相比，DPI 不仅对原来的 TCP/IP 4 层（IP 层）以下的协议进行分析，还增加了 4 层以上信息的检测，因而其提供的信息更丰富，但处理也更为复杂。

传统的 DPI 检测系统通过将报文负荷和预先设定的字符串集合进行比较来判断报文是否满足特定特征。近年来，越来越多的系统采用正则表达式取代字符串来描述报文特征。和字符串相比，正则表达式能够非常灵活、简单、有效地描述各种特征，使得特征串具有动态特性，适合各种动态搜索，比如 b,ab,aab,aaab,aaaab…这一系列的字符串特征可以简单地用一个正则表达式 a*b 来表示。不同的正则表达式语言规范的描述能力不同，一般目前业界比较流行的正则表达式规范有可移植的操作系统界面（Portable Operating System

Interface，POSIX）和 Perl 兼容的正则表达式（Perl Compatible Regular Expression，PCRE）两种。PCRE 中增加了一些 POSIX 不支持的扩展，具有更强大的表达能力。比如开源代码 Snort 就采用 PCRE 规范来描述其部分规则。目前大部份设备还是使用 POSIX 规范，有些设备号称能够支持 PCRE 规范，但是从其产品资料分析，其实仍然只是 PCRE 规范中兼容 POSIX 规范的一个子集，并没有真正支持复杂的 PCRE 语法。

判断报文内容中是否包含正则表达式表示的规则的操作称为正则表达式匹配，目前，业界比较流行的正则表达式匹配方法有:

一种方法是基于确定性有限状态机DFA方法，这类方法都是预先将正则表达式转换为某种形式描述的状态转换表，匹配过程中，将报文中的符号作为输入条件查询状态转换表确定下一个跳转状态来实现状态迁移。这种方法的优点是“查表－跳转”的操作模式比较简单，能够很方便用硬件实现，匹配速度很快。缺点是只支持较为简单的正则表达式规范，对PCRE中的很多扩展选项（比如条件表达式、^、$等和位置相关的符号、匹配选项）都无法很好地支持。此外，部分正则表达式（比如：.*AB.{j}CD）用DFA表示时，其状态数目随正则表达式的长度呈指数级增长关系，给存储带来很大的负担。

另一种方法是非确定性有限状态机NFA方法，其基本原理类似于DFA方法，所不同的是：NFA允许出现空符号（即不输入任何符号实现跳转），且NFA输入一个符号可以对应多个下一个跳转状态。这种不确定性导致其用硬件实现非常困难，目前也有一些使用硬件实现NFA的方法，一般都是将NFA状态表直接用逻辑器件实现，这样如果更新正则表达式需要对逻辑器件更新，存在扩展性差的问题。

第三种方法是程序解析的方法。这类方法一般不显式生成状态转换表。以 PCRE 源码库为例，它将正则表达式分解为程序可以理解的最小片段。匹配过程中，程序将报文中的符号按照出现的位置落入不同的片段中，如果符号能够符合这个片段，则要么在这个片段中等待下一个符号，或者进入下一

个片段。一些软件实现可能采用其他特定的处理方式，但都可以归类为程序解析。这种方法通常具有很良好的扩展性，能够支持较为复杂的正则表达式语法，但是和状态机的方式比，它的匹配速度比较慢，容易成为整个系统的瓶颈。

现有技术一给出一种分层的报文过滤架构，在每一层过滤器上可以定义不同的过滤标准。通过第一层过滤的报文送往第二层过滤器进行过滤，直到得出过滤结果。现有技术一适合于各层完成的过滤标准不相同的情况，而且通常由较先出现的层次完成一定的筛选，减少后面层次所需的数据处理量。因而各层次之间有严格的先后顺序，从提高整体效率的角度考虑，应该将处理越复杂的过滤器越往后放置。现有技术一主要有以下缺点：各层过滤器间有严格的顺序关系，不够灵活；会造成报文的处理延迟时间加长，比如一些报文根本不需要第一层过滤器进行处理，但是按照层次化过滤的架构，这些报文也需要经过第一层过滤器后再进入第二层过滤器进行匹配，一方面浪费了第一层过滤器的处理能力，另一方面也造成了整个报文匹配的时间增加。只是笼统地给出了过滤标准的定义，并没有给出针对正则表达式匹配的具体方案;

请参阅图1，为现有技术二利用三重内容可寻址存储器（Ternary Content Addressable Memory，TCAM）304存储DFA来实现正则表达式匹配系统的结构示意图。利用TCAM304存储DFA来实现正则表达式匹配的方法。其实现步骤可以描述为:

1. 利用正则表达式中的元字符（如.*）将一个正则表达式分为多个子表达式，使用 TCAM304 存储正则表达式（实际存放的是 DFA 状态表）。利用 TCAM304 可以一次比较多个字符的特性，可以以多步长方式实现状态机;
2. 在第二内存结构 320 中存放报文处理动作，该第二内存结构 320 可以为 RAM;
3. 预解析器，即图中前置分析程序 334 从接受的包缓存 258 中提取需要处理

的字段放在消息缓存中;

4. 消息缓存区 306 用来存放需要进行正则表达式匹配的报文消息;
5. 解码电路 302 对报文解码，执行和报文相关的指令;
6. 在解码电路 302 的控制下，桶式移位器 308 将报文中的不同部分取出，和标签空间（Tag Space）318 中存放的当前状态一起送到 TCAM304 中进行比较;
7. 一旦 TCAM304 检测出正则表达式，则第二内存结构 320 中对应的动作会被译码电路执行，译码电路产生信号给流量控制器 352。

现有技术二主要有以下缺点:现有技术二中技术方案的实现和TCAM可以同时访问多位的特性密切相关，不具有通用性；TCAM价格昂贵、功耗较大也限制了其使用;方案中为了减少子表达式的数目以达到减少占用TCAM条目数的目的，需要将不同的DFA合成为一个DFA，如果对正则表达式进行更新，需要重新编译整个DFA；对于PCRE中的很多扩展选项都无法很好地支持。

现有技术三中，在整个检测系统中使用专门的正则表达式检测模块，遇到需要检测正则表达式时将流量输入到该模块进行处理，该模块采用ASIC芯片实现DFA,目前只支持POSIX正则表达式,实现细节属于商业秘密无法获知。现有技术三的缺点：无法支持PCRE正则表达式；经分析得知，如果需要支持PCRE正则表达式，需要用户在ASIC芯片以外提供实现，且没有给出具体的实现方案。

发明内容

本发明的实施例的目的在于提供一种正则表达式编译、匹配系统及编译、匹配方法，旨在解决现有技术中的系统没有考虑到正则表达式本身的特点，仅采用 DFA 方法处理所有正则表达式，无法真正支持 PCRE 正则表达式的问题。

本发明的实施例是这样实现的，一种正则表达式编译系统，包括词法分析模块和至少两种编译模块，所述词法分析模块分析正则表达式的词法特点，

根据预设词法规则和所述正则表达式的词法特点将所述正则表达式送往相应的的编译模块处理，所述相应的编译模块接收正则表达式，将所述正则表达式编译成特定形式的数据结构。

本发明的实施例所采取的另一技术方案包括一种正则表达式匹配系统，包括至少两种匹配模块，所述匹配模块按照接收到的数据报文所符合的正则表达式的词法特点，对该数据报文进行相应的匹配，所述匹配方式包括 DFA 匹配和 PCRE 匹配。

本发明的实施例所采取的又一技术方案包括一种正则表达式编译方法，包括:

接收用户输入的正则表达式规则，对正则表达式进行词法解析，其中所述正则表达式规则包括所述正则表达式;

根据预设规则将所述正则表达式编译成特定形式的数据结构。

本发明的实施例所采取的又一技术方案包括一种正则表达式匹配方法，包括:

接收前端模块发送的需要进行匹配操作的数据报文;

按照所述数据报文符合的正则表达式的词法特点，对所述数据报文进行相应的匹配。

本发明的实施例的正则表达式匹配系统及匹配方法通过结合使用多种正则表达式匹配算法，既保证了正则表达式匹配引擎具有较高的处理速度，又保证了对正则表达式的各种扩展语法的兼容性；本发明的实施例的正则表达式编译系统及编译方法通过使用具有反馈机制的编译器，能够根据实时流量的特征调整引擎的处理行为，提高系统吞吐量。

本发明的特征及优点将通过实施例结合附图进行详细说明。

附图说明

图 1 是现有技术二利用 TCAM 存储 DFA 来实现正则表达式匹配系统的结构示意图;

图 2 是本发明实施例的正则表达式匹配引擎的结构示意图;

图 3 是本发明实施例的正则表达式匹配系统采用单通道链接方式的结构示意图;

图 4 是本发明实施例的正则表达式匹配系统采用并联链接方式的结构示意图;

图 5 是本发明实施例的正则表达式匹配系统采用串联链接方式的结构示意图。

具体实施方式

为了使本发明实施例的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明实施例进行进一步详细说明。

请参阅图 2，为本发明实施例的正则表达式匹配引擎的结构示意图。本发明实施例的正则表达式匹配引擎包括正则表达式编译系统和正则表达式匹配系统。正则表达式编译系统与系统接口相连，接收用户通过该系统接口输入的需要匹配的正则表达式规则，该正则表达式规则中包含正则表达式。正则表达式匹配系统与前端模块连接，接收需要进行匹配操作的数据报文，正则表达式匹配系统还可以接收正则表达式信息，用来指示需要匹配的正则表达式。

本发明实施例的正则表达式编译系统包括词法分析模块、分类信息库、DFA 编译模块、DFA 状态转换表、PCRE 编译模块和 PCRE 片段。词法分析模块接收到用户输入的正则表达式规则，对其中的正则表达式进行词法解析，根据预设规则和正则表达式的词法特点确定该正则表达式是送到 DFA 编译模块，还是 PCRE 编译模块进行编译，该预设规则是对正则表达式进行分类的规则,包括根据 DFA 词法特点设置的规则和根据 PCRE 词法特点设置的规则;同时，由上述确定的结果生成决策信息，将该决策信息被记录在分类信息库中。如果 DFA 编译模块接收到正则表达式，将正则表达式编译为 DFA 状态转换表，并存储，如果 PCRE 编译模块接收到正则表达式，将正则表达式编

译为 PCRE 片段，并存储。

本发明实施例的正则表达式的编译方法包括:

用户通过系统接口（比如命令行或操作界面等）将需要匹配的正则表达式输入到引擎;

词法分析模块接收到用户输入的正则表达式，对正则表达式进行词法解析，根据预设规则和正则表达式的词法特点确定该正则表达式是送到 DFA 编译模块，还是 PCRE 编译模块进行编译。同时，由上述确定的结果生成决策信息，该决策信息被记录在分类信息库中;

如果 DFA 编译模块接收到正则表达式，将正则表达式编译为 DFA 状态转换表，并存储，其中，DFA 状态转换表并不一定指原始的二维转换表或转换图等特定形式，出于其他因素如空间效率考虑，还可能对状态转换表进行压缩，对不同的 DFA 进行合并等操作，这时状态转换表的存储形式可能发生变化，不影响本发明的实质内容;

如果 PCRE 编译模块接收到正则表达式，将正则表达式编译为 PCRE 片段，并存储。在实际软件实现时 PCRE 片段的存储形式是和具体实现相关的，形式上的不同不影响本发明的实质内容;

本发明实施例的正则表达式匹配系统包括仲裁模块、DFA 适配器、PCRE 适配器和综合输出模块。前端模块发送需要进行匹配操作的数据报文和正则表达式信息（可选）给引擎，其中，数据报文可以是通常所说的 IP 报文，为方便起见，前端的某个模块，包括一个或多个模块，在图 2 中示例性的表示为前端模块，可以将原始 IP 报文中的三层、四层协议头移除后交给正则表达式匹配引擎处理；为了跨报文检测某个正则表达式，上述前端模块还可能将报文进行重组；对于需要综合输出模块进行排序、统计的情况，需要送入的数据报文中有用于标识报文的字段，比如原始报文中的顺序号或者由前端模块分配的某种标识，不同的报文格式不影响本发明实施例的实质内容。图中前端模块送给仲裁模块的正则表达式的标识符与词法分析模块为正则表达式

设定的标识符相对应，用来指明需要检测的输入数据报文中含有哪个正则表达式。该标识符在编译阶段产生，由前端模块根据自己的处理结果和某种映射规则确定该标识符。这样做的好处是在不同的正则表达式用不同 DFA 表示时，可以缩小需要匹配的正则表达式范围；如果由 DFA 编译模块编译的所有正则表达式被编译进一个大 DFA 中，则在进行匹配操作时，前端模块不需要指定标识符；不同 DFA 实现方式下入口参数的变化应不影响本发明实施例的实质内容。仲裁模块查找正则表达式相关的分类信息，该分类信息中可以记录正则表达式已经被编译为 DFA 还是 PCRE，通过查找分类信息确定由哪个适配器进行匹配处理，在进行匹配操作时，仲裁模块可以确定数据报文的流向和顺序，例如，仲裁模块通过分类信息中携带的数据报文的流向和顺序信息，确定数据报文的流向和顺序；可以只由一个 DFA 适配器或 PCRE 适配器进行处理，也可以由多个 DFA 适配器和/或 PCRE 适配器进行处理。

本发明实施例的正则表达式匹配方法包括:

前端模块发送需要进行匹配操作的数据报文和正则表达式信息（可选）给引擎;

仲裁模块查找正则表达式相关的分类信息，根据查找结果确定由哪个适配器进行匹配处理，当正则表达式被编译为 DFA 时，将该正则表达式送往 DFA 适配器，当正则表达式被编译为 PCRE，将该该正则表达式送往 PCRE 适配器;

DFA 适配器/PCRE 适配器对数据报文进行适配，将适配/匹配结果（至少包括匹配成功信息和匹配失败信息）统一送到综合输出模块，如果有多个 DFA/PCRE 适配器线程对不同的数据报文进行适配/匹配处理，可能因为数据报文的长度不同、需要匹配的正则表达式的复杂程度不同等因素，造成先收到的报文比后收到的报文的处理时间长，这时如果直接输出匹配结果给后续业务处理模块，匹配结果可能是乱序的。如果这种顺序很重要，可以在综合输出模块中设置一个缓存来对报文匹配结果进行排序；排序的方法有多种，

比如可以根据报文的序号来进行。可选的，该综合输出模块还可以对匹配结果进行数据统计，数据统计条件可以灵活设定，比如在最近 30 分钟内被 DFA 适配器处理的前 10 位的正则表达式；不同的统计条件不影响本发明的实质内容；

综合输出模块将匹配结果送给后续业务处理模块，作为后续业务处理模块执行动作的依据，可选的，综合输出模块可以将统计信息反馈给词法分析模块，作为其分发正则表达式的参考依据，进行二次编译，比如当检测到某个正则表达式最近经常由 DFA 适配器处理，词法分析模块可以将此正则表达式再送到 PCRE 编译模块进行编译，当 DFA 适配器忙不过来时，仲裁模块可以协调该报文到 PCRE 适配器进行处理。

在进行匹配操作时，仲裁模块可以确定数据报文的流向和顺序，例如，仲裁模块通过分类信息中携带的数据报文的流向和顺序信息，确定数据报文的流向和顺序；可以只由一个 DFA 适配器或 PCRE 适配器进行处理，也可以由多个 DFA 适配器和/或 PCRE 适配器进行处理；按照各适配器间的链接关系不同，可以分为单通道、并联和串联三种链接方式。

请参阅图3，单通道链接方式是仲裁模块将数据报文调度到DFA适配器或者是PCRE适配器进行处理，数据只流向唯一的适配器。为了简化起见，在以下的图中都省略了控制层面的模块（有些情况下还省略了综合输出模块）。如图3(a)所示，大部分情况下，一条用户输入的正则表达式规则中只包含某种特定的正则表达式，该正则表达式符合DFA或PCRE的词法特点，由仲裁模块将报文送到DFA或PCRE适配器进行处理；另外一种实施方式如图3(b)所示，如果某个数据报文应该是由DFA适配器进行处理，但DFA适配器处理不过来，而这时PCRE适配器刚好空闲，可以由仲裁模块将该数据报文送到PCRE适配器进行处理。这种情况下两种适配器可以是用不同的器件实现，比如DFA适配器用FPGA实现，PCRE适配器用多核实现。这样处理能很方便地实现负载均衡，提高整个引擎的吞吐量。

请参阅图4，并联链接方式和单通道方式的实施方式如图3(b)的模块关系很类似，不同的是，数据报文会被分配到各个适配器进行处理。有一些用户输入的正则表达式规则中包含有多个正则表达式，部分正则表达式适合用DFA方式表示，另外一部分正则表达式只能用PCRE方式表示。当有数据报文需要匹配这样的规则时，两个适配器可以同时处理报文。这种方式的优点是缩短了整个正则表达式匹配引擎的处理时间。

请参阅图5，串联链接方式下仲裁模块可以单独存在，链接在DFA适配器前；也可以不需要实现仲裁模块，将数据报文和正则表达式缺省先送到DFA适配器，隐含地由DFA适配器兼任仲裁模块的角色。如图5(a)所示，在一条用户规则中既存在DFA方式表示的正则表达式，也存在PCRE方式表示的正则表达式的情况下，先通过DFA适配器过滤掉不满足DFA方式表示的正则表达式的数据报文，可以避免后续不必要的PCRE匹配操作；或者如图5 (b)所示，数据报文缺省被送到DFA适配器，如果DFA适配器能按照DFA词法特点对该数据报文进行匹配能够处理，就执行DFA匹配操作，输出结果；否则，将数据报文再送到PCRE适配器进行处理。这样做的优点是模块间结构较简单，不需要单独实现仲裁模块。

本发明实施例的正则表达式编译、匹配系统及编译、匹配方法以 DFA 和 PCRE 适配器作为实例进行说明，本发明的实施例也适合采用其他的匹配方法来处理正则表达式，比如采用 DFA 和 NFA 方法。本发明主要是保护一种结合使用多种匹配算法进行正则表达式匹配的系统和方法，对使用的算法并不做限制。本发明实施例的正则表达式编译、匹配系统及编译、匹配方法只采用了两种编译/匹配模块，但是可以很方便地扩展到更多种类的编译/匹配模块；模块的数量不影响本发明的实质内容。

本发明实施例中，只列举了一个 DFA 适配器和一个 PCRE 适配器；在实际实现时，没有这个限制；事实上为了达到最大的处理性能，往往会同时使用多个 DFA 适配器和多个 PCRE 适配器，比如采用多个线程。此时，仲裁模

块可以根据各线程的负荷情况对报文进行分流。这不影响本发明的实质内容。

本发明以 PCRE 规范的正则表达式为实例进行说明，但是方法对于其他可能的正则表达式规范也适用。采用不同的正则表达式规范不影响本发明的实质内容。

本发明实施例的正则表达式编译、匹配系统及编译、匹配方法是针对网络环境下对数据报文进行高速正则表达式匹配的情况，但对于搜索引擎、数据库检索、自然语言理解、文档分类等这些情况下可能使用其他的正则表达式规范，而且正则表达式匹配对象也不局限于报文，比如数据库检索时针对的是数据表中的数据条目，但本发明提供的方法和系统同样适用。

本发明实施例的匹配系统可以在不同的器件中实现，也可以在相同器件中实现；可以使用硬件的方式实现，也可以采用软件的方式实现，甚至可以作为同一个软件中不同的函数来实现。因此本发明具有实现方式灵活的特点，但实现方式不同不影响本发明的实质内容。

本发明实施例的正则表达式编译、匹配系统及编译、匹配方法考虑了正则表达式本身的特点，能真正支持PCRE正则表达式，可以有效减少设备的存储器件成本；在对正则表达式进行编译时，如果正则表达式集合不经常变化，可以选择将所有正则表达式编译到同一个DFA以节省存储空间；在正则表达式集合需要经常更新的情况下，将不同的正则表达式编译到不同DFA，可以在更新个别正则表达式时，不需要对整个DFA进行重新编译，节省编译时间；不管在以上哪种情况下，由于词法分析已经将状态数目较多的正则表达式区分出来，交由PCRE编译器处理，因此可以节省存储空间；本发明实施例的正则表达式编译、匹配系统及编译、匹配方法在保证网络设备报文高速处理的同时，提供对类似于PCRE的正则表达式规范的全面支持。

本发明有益的技术效果在于：在完全支持PCRE正则表达式语法的前提下，达到较高的处理速度；能够根据实时流量的特征调整引擎的处理行为，提高系统吞吐量。本发明实施例可应用于数据报文处理、搜索引擎、数据库

检索、自然语言理解、文档分类等可能使用正则表达式的领域。

以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1、一种正则表达式编译系统，其特征在于，包括词法分析模块和至少两种编译模块，所述词法分析模块分析正则表达式的词法特点，根据预设词法规则和所述正则表达式的词法特点将所述正则表达式送往相应的编译模块，所述相应的编译模块接收所述正则表达式，将所述正则表达式编译成特定形式的数据结构。

2、如权利要求 1 所述的正则表达式编译系统，其特征在于，还包括分类信息库，所述词法分析模块产生的决策信息记录在分类信息库中。

3、如权利要求 2 所述的正则表达式编译系统，其特征在于，所述编译模块包括确定性有限状态机 DFA 编译模块和 Perl 兼容的正则表达式 PCRE 编译模块，所述 DFA 编译模块收到正则表达式，将正则表达式编译为 DFA 状态转换表并存储，所述 PCRE 编译模块接收到正则表达式，将正则表达式编译为 PCRE 片段，并存储。

4、如权利要求 2 所述的正则表达式编译系统，其特征在于，所述编译模块包括确定性有限状态机 DFA 编译模块和非确定性状态机 NFA 编译模块，所述 DFA 编译模块收到正则表达式，将正则表达式编译为 DFA 状态转换表并存储，所述 NFA 编译模块接收到正则表达式，将正则表达式编译为 NFA 状态转换表并存储 。

5、一种正则表达式匹配系统，其特征在于，包括至少两种匹配模块，所述匹配模块按照接收到的数据报文所符合的正则表达式的词法特点，对该数据报文进行相应的匹配，所述匹配方式包括 DFA 匹配和 PCRE 匹配。

6、如权利要求 5 所述的正则表达式匹配系统，其特征在于，还包括仲裁模块，所述仲裁模块根据预设规则确定数据报文在各匹配模块间的处理顺序，所述匹配模块间的链接方式包括单通道链接方式、并联链接方式和串联链接方式。

7、如权利要求 5 或 6 所述的正则表达式匹配系统，其特征在于，还包括综

合输出模块，所述综合输出模块与匹配模块相连接，接收匹配模块输出的匹配结果。

8、如权利要求 7 所述的正则表达式匹配系统，其特征在于，所述综合输出模块中设置有缓存，用来对报文匹配结果进行排序。

9、如权利要求 7 所述的正则表达式匹配系统，其特征在于，所述综合输出模块对所述匹配结果进行数据统计，获取统计信息。

10、如权利要求 9 所述的正则表达式匹配系统，其特征在于，所述综合输出模块将所述统计信息反馈给词法分析模块，由词法分析模块根据所述统计信息更新预设词法规则，并发起二次编译。

11、一种正则表达式编译方法，其特征在于，该方法包括:

接收用户输入的正则表达式规则，对正则表达式进行词法解析，其中所述正则表达式规则包括所述正则表达式;

根据预设规则将所述正则表达式编译成特定形式的数据结构。

12、如权利要求 11 所述的正则表达式编译方法，其特征在于，所述正则表达式规则通过词法分析模块进行接收。

13、如权利要求 11 所述的正则表达式编译方法，其特征在于，所述的根据预设规则将正则表达式编译成特定形式的数据结构包括:

DFA 编译模块收到正则表达式，将正则表达式编译为 DFA 状态转换表并存储；或者，

PCRE 编译模块接收到正则表达式，将正则表达式编译为 PCRE 片段并存储。

14、一种正则表达式匹配方法，其特征在于，该方法包括:

接收前端模块发送的需要进行匹配操作的数据报文;

按照所述数据报文符合的正则表达式的词法特点，对所述数据报文进行相应的匹配。

15、如权利要求 14 所述的正则表达式匹配方法，其特征在于，所述方法还

包括：根据预设规则确定数据报文在各匹配模块间的处理顺序，所述匹配模块间的处理方式包括单通道链接方式、并联链接方式和串联链接方式。

16、如权利要求 15 所述的正则表达式匹配方法，其特征在于，所述匹配模块将匹配结果统一送到综合输出模块。

17、如权利要求 16 所述的正则表达式匹配方法，其特征在于，所述综合输出模块中设置有缓存，用来对报文匹配结果进行排序。

18、如权利要求 16 所述的正则表达式匹配方法，其特征在于，所述综合输出模块对匹配结果进行数据统计，获取统计信息。

19、如权利要求 18 所述的正则表达式匹配方法，其特征在于，所述综合输出模块将所述统计信息反馈给词法分析模块，由词法分析模块根据所述统计信息更新预设词法规则，并发起二次编译。

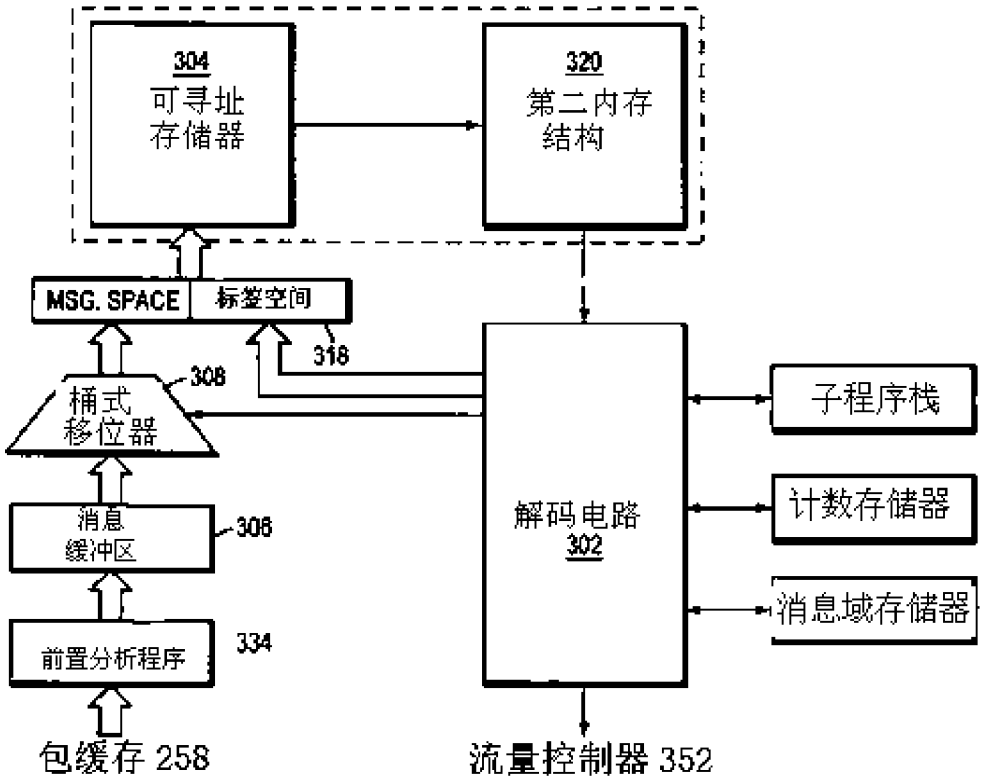
304
可寻址
存储器
320
第二内存
结构
MSG. SPACE
标签空间
318
308
桶式
移位器
消息
缓冲区
306
前置分析程序
334
包缓存 258
解码电路
302
流量控制器 352
子程序栈
计数存储器
消息域存储器

图 1

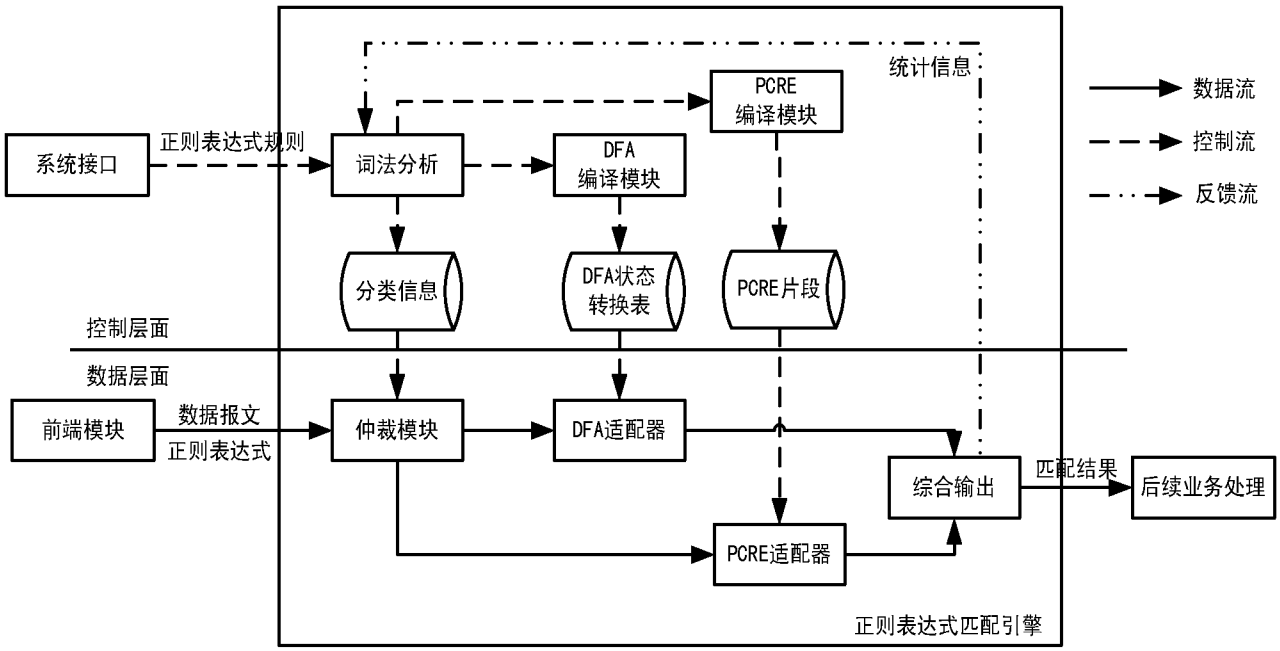
统计信息
PCRE
编译模块
数据流
控制流
反馈流
系统接口
正则表达式规则
词法分析
DFA
编译模块
分类信息
DFA状态
转换表
PCRE片段
控制层面
数据层面
前端模块
数据报文
正则表达式
仲裁模块
DFA适配器
综合输出
匹配结果
后续业务处理
PCRE适配器
正则表达式匹配引擎

图 2

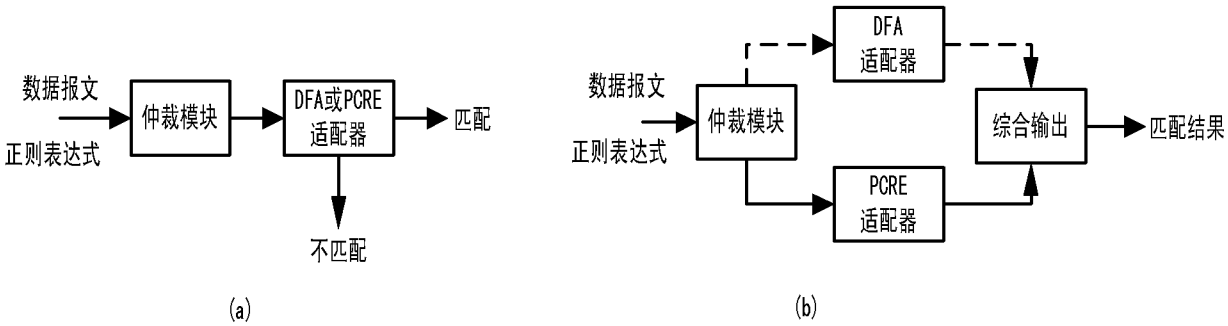
数据报文
正则表达式
仲裁模块
DFA或PCRE
适配器
匹配
不匹配
(a)
数据报文
正则表达式
仲裁模块
DFA
适配器
PCRE
适配器
综合输出
匹配结果
(b)

图 3

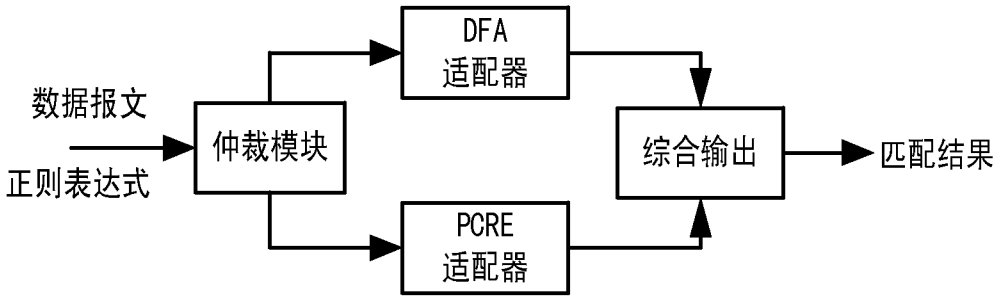
数据报文
正则表达式
仲裁模块
DFA
适配器
PCRE
适配器
综合输出
匹配结果

图 4

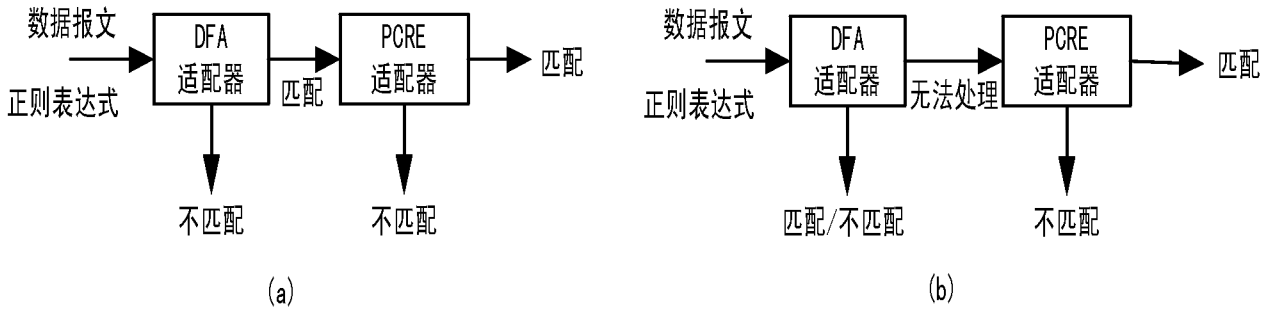
数据报文
正则表达式
DFA
适配器
匹配
PCRE
适配器
匹配
不匹配
不匹配
(a)
数据报文
正则表达式
DFA
适配器
无法处理
PCRE
适配器
匹配
匹配/不匹配
不匹配
(b)

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2008/071811

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/30(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC:G06F;H04L;H04Q;H04B;H04M

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI;EPODOC;PAJ;CNKI;IEEE;CNPAT: regular,expression,compil+,match+,pars+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP2007102744A (MAEDA Y) 19 Apr.2007 (19.04.2007) paragraphs [0035]-[0080] and figure 1 in description	11-13
X	US2007/0055664A1 (CISCO TECHNOLOGY,INC.,) 08 Mar.2007 (08.03.2007) paragraphs [0017]-[0074] and figures 1-8 in description	5-10,14-19
A	CN1255674A (SUN MICROSYSTEMS,INC.,) 07 Jun.2000 (07.06.2000) the whole document	1-19
A	US7225188B1 (CISCO TECHNOLOGY,INC.,) 29 May 2007 (29.05.2007) the whole document	1-19

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim (S) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
02 Sep.2008(02.09.2008)	**18 Sep. 2008 (18.09.2008)**
Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer ZHANG,Renjie Telephone No. (86-10)62413133

Form PCT/ISA/210 (second sheet) (April 2007)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2008/071811

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
JP2007102744A	19.04.2007	NONE	
US2007/0055664A1	08.03.2007	NONE	
CN1255674A	07.06.2000	EP0997816A2	03.05.2000
		JP2000181724A	30.06.2000
		CA2287746A1	30.04.2000
		US6298477B1	02.10.2001
US7225188B1	29.05.2007	NONE	

国际检索报告

国际申请号
PCT/CN2008/071811

A. 主题的分类

G06F17/30(2006.01)i

按照国际专利分类表(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC:G06F;H04L;H04Q;H04B;H04M

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

WPI;EPODOC;PAJ;CNKI;IEEE;CNPAT:正则，正规，表达式，编译，匹配，解析,regular,expression,compil+,match+,pars+

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
X	JP 特开 2007-102744A（前田泰成）19.4 月 2007（19.04.2007）说明书第[0035]-[0080]段，图 1	11-13
X	US2007/0055664A1（CISCO TECHNOLOGY,INC.,）08.3 月 2007（08.03.2007）说明书第[0017]-[0074]段，图 1-8	5-10,14-19
A	CN1255674A（太阳微系统有限公司）07.6 月 2000（07.06.2000）全文	1-19
A	US7225188B1（CISCO TECHNOLOGY,INC.,）29.5 月 2007（29.05.2007）全文	1-19

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
02.9 月 2008（02.09.2008）	**18.9 月 2008 (18.09.2008)**
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	受权官员 张仁杰 电话号码：(86-10) **62413133**

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2008/071811

检索报告中引用的专利文件	公布日期	同族专利	公布日期
JP 特开 2007-102744A	19.04.2007	无	
US2007/0055664A1	08.03.2007	无	
CN1255674A	07.06.2000	EP0997816A2	03.05.2000
		JP2000181724A	30.06.2000
		CA2287746A1	30.04.2000
		US6298477B1	02.10.2001
US7225188B1	29.05.2007	无	