计算机应用 2013, 33(08) 2370-2374 DOI: ISSN: 1001-9081 CN: 51-1307/TP

典型应用

基于多字符DFA的高速正则表达式匹配算法

贺炜,郭云飞,莫涵,扈红超

国家数字交换系统工程技术研究中心，郑州 450002

摘要： 基于确定性有限自动机(DFA)的传统正则表达式匹配方法存在单周期处理单字符的速度瓶颈。为提升处理速率，提出一种单周期处理多字符的匹配算法MC-DFA，该算法基于DFA实现，支持匹配位置的精确定位。MC-DFA将传统DFA中的单字符跳转合并为多字符跳转，实现了单周期处理多个输入字符。通过状态转移矩阵二阶压缩算法，MC-DFA分别对矩阵行内以及行间冗余进行消除，减少了内存使用。300条规则下，单周期处理8字符时，MC-DFA吞吐率能够达到7.88Gb/s，内存占用小于6MB，预处理时间为19.24s。实验结果表明，MC-DFA能够有效提升系统吞吐率，并且保证内存占用在可接受范围之内，性能优于现有正则表达式匹配算法。

关键词： 正则表达式 高速 多字符 精确定位 矩阵压缩

Multi-character DFA-based high speed regular expression matching algorithm

HE Wei,GUO Yunfei,MO Han,HU Hongchao

China National Digital Switching System Engineering and Technological R&D Center, Zhengzhou Henan 450002, China

Abstract: Traditional Deterministic Finite Automata (DFA) based regular expression matching can only process one character per cycle, which is the speed bottleneck. A new algorithm named Multi-Character DFA (MC-DFA) was proposed for high throughput matching and precise positioning. It combined the one character transition in traditional DFA together to handle multi-character processing per cycle. A new transition matrix compress algorithm was also proposed to reduce the redundancy introduced by MC-DFA. The result demonstrates that MC-DFA can improve the throughput efficiently while requiring acceptable memory. For a set of 300 regexes, 8C-DFA obtains a throughput of 7.88Gb/s, memory usage less than 6MB and 19.24s preprocessing time, better than traditional methods.

Keywords: regular expression high throughput multi-character precise positioning matrix compress

收稿日期 2013-02-20 修回日期 2013-03-12 网络版发布日期 2013-09-11

DOI:

基金项目:

国家科技支撑计划项目;国家863计划项目

通讯作者: 贺炜

作者简介:

贺炜(1988-)，男，山西吕梁人，硕士研究生，主要研究方向：网络安全中的特征匹配;
郭云飞(1963-)，男，河南郑州人，教授，博士生导师，主要研究方向：三网融合;
莫涵(1988-)，女，河南郑州人，硕士研究生，主要研究方向：可重构网络、组播传输;
扈红超(1983-)，男，河南郑州人，研究员，博士，主要研究方向：网络协议分析。
作者Email: hewxiaoyao@163.com

参考文献：

本刊中的类似文章

1．罗强 王倩 刘方林 范瑞娟.基于SCA/SDO的高速铁路综合调度系统中间件设计[J]. 计算机应用, 2013,33(06): 1654-1669

2．李娟.形式语言在网页制作操作题自动阅卷中的应用[J]. 计算机应用, 2013,33(03): 882-885

3．郑争兵.基于FPGA的高速采样缓存系统的设计与实现[J]. 计算机应用, 2012,32(11): 3259-3261

4．蒋新华 朱铨 邹复民.高速铁路3G通信的覆盖与切换技术综述[J]. 计算机应用, 2012,32(09): 2385-2390

5．张墨华 李戈.基于中间点划分无冲突哈希的高速包处理[J]. 计算机应用, 2012,32(04): 999-1002

6．张其亮 陈永生 杜磊.基于编织算法的复线高速磁浮列车运行图铺画方法[J]. 计算机应用, 2011,31(12): 3434-3437

7．唐球 姜磊 谭建龙 刘金刚.基于FPGA的正则表达式匹配算法综述[J]. 计算机应用, 2011,31(11): 2943-2946

8．赵文波 孙小科 马草川.基于非线性窗口增长的TCP Westwood改进算法[J]. 计算机应用, 2011,31(09): 2344-2348

9．李敏.高分辨率合成孔径雷达图像高速公路检测法[J]. 计算机应用, 2011,31(07): 1825-1826

10．刘洪涛 程良伦.基于优先级的服务区分和速率控制策略[J]. 计算机应用, 2011,31(06): 1458-1460

11．姚远 刘鹏 王辉 笱程成.基于稀疏矩阵存储的状态表压缩算法[J]. 计算机应用, 2010,30(8): 2157-2160

12．熊静 喻钢 徐中伟 郦萌.面向安全评估的高速铁路CTCS-2列控系统安全性测试环境[J]. 计算机应用, 2010,30(8): 2181-2184

13．张华 胡修林.基于PCI9656的高速实时采集存储系统[J]. 计算机应用, 2010,30(11): 3130-3133

14．李岩 崔晓英 李贤尧 赵宏杰 程平.μC/OS-Ⅱ任务管理的硬件实现[J]. 计算机应用, 2010,30(05): 1386-1389

15．傅志中 鲜海滢 陈友林.基于PCI总线的高速数据采集设备驱动开发[J]. 计算机应用, 2009,29(2): 577-579

16．姚远 刘鹏 单征 田双鹏.面向存储的正则表达式匹配算法综述[J]. 计算机应用, 2009,29(12): 3171-3173

17．张加万 施翠翠.基于多核计算平台和高速缓存感知的Haar小波变换算法[J]. 计算机应用, 2009,29(08): 2139-2142

18．杨仲唱 张信明.基于飞思卡尔i.MX31的Standalone开发平台设计[J]. 计算机应用, 2008,28(4): 1052-1054

19．丁晶 陈晓岚 吴萍.基于正则表达式的深度包检测算法[J]. 计算机应用, 2007,27(9): 2184-2186

20．张毅坤 刘伟.基于自动机模型的构件集成软件测试要素的提取[J]. 计算机应用, 2007,27(4): 857-859

21．吴亮 高建强 穆建成.客运专线的追踪间隔控制模型与计算[J]. 计算机应用, 2007,27(11): 2643-2645

22．石磊 姚瑶.马尔可夫预测模型的压缩与应用研究[J]. 计算机应用, 2007,27(11): 2746-2749

23．徐春 林忠钦 李淑慧 夏年炯 .基于Hough变换的双球冲球心精确定位技术研究[J]. 计算机应用, 2006,26(9): 2051-2053

24．曾华燊 高雨 .论高速接入网技术[J]. 计算机应用, 2006,26(8): 1751-1755

25．满红芳.高速环境下基于数据分流的入侵检测系统设计[J]. 计算机应用, 2005,25(12): 2734-2735
