文章编号:1008-0570(2010)02-3-0091-02

基于值域二次哈希方法的布鲁姆过滤器

Bloom Filter Based on Value Hashing Method

(1.湖南大学;2.湖南民族职业学院;3.东莞理工学院) 张生华[1] 秦 拯[1] 宋 勇[2] 张忠志[3]
ZHANG Sheng-hua QIN Zheng SONG Yong ZHANG Zhong-zhi

摘要: 本文针对扩展式布鲁姆过滤器(EBF)内存消耗过大,提出一种基于值域哈希二次过滤的布鲁姆过滤器数据结构(VHBF)和相关算法,VHBF 通过在布鲁姆过滤器中对集合中的每个特征进行 k 次哈希,并将此 k 次哈希值转化为相应特征的镜像特征。然后对此镜像进行二次过滤运算,运算后的结果保存在另一布鲁姆过滤器中。在对特征进行检索时,由于无需保存特征本身,因而空间效率比 EBF 更高。实验表明,VHBF 的假阳性误判率的比扩展型布鲁姆过滤器(EBF)低,而 VHBF 内存消耗也低于 EBF。

关键词: 特征检测; 布鲁姆过滤器; 哈希; 成员查找
中图分类号: TP393 **文献标识码:** A

Abstract: This paper proposes efficient data structure called Value Hash Bloom Filter (VHBF) and the corresponding algorithms. In the programming stage of VHBF, the hash results of the first clusters of hash functions in the first Bloom Filter will be connected as a mirror image of the inserted signature. This mirror image will be hashed into another bloom filter like array. And in VHBF, it is not necessary to store all the actual signatures. Thus, with the mirror image information, we are then able to reduce the total consumption of memory involved in the membership query .Analysis and experiments show that the Value Hash Bloom Filter is significantly efficient for practical purposes than the Extended Bloom Filter and improved Extended Bloom Filter.

Key words: signature detection; bloom filter; hash; member query

技术创新

引言

随着网络信息复杂化, 防火墙和网络监测系统以及网络流量特征分析与测量急需高效率的特征检测算法。

入侵检测系统如 snort 通过识别可疑特征阻止对系统的攻击与破坏的企图。入侵检测系统中最常使用的方法是异常统计检测与模式匹配检测。研究显示,特征匹配模块是提高 snort 性能的关键点。本文描述了一种适合使用在高速入侵检测系统中的特征匹配引擎。我们使用一种基于 Bloom Filter 的特征匹配引擎,以改进现有的特征匹配技术,提高系统在高速网络应用上的吞吐量。

1 布鲁姆过滤器

布鲁姆过滤器(Bloom Filter)是一种空间效率很高的随机数据结构,它利用位数组很简洁地表示一个集合,并能判断一个元素是否属于这个集合。布鲁姆过滤器结构的实质是将集合中的元素通过 k 个 Hash 函数映射到位串向量中,对于一个元素只需要保存几个比特。布鲁姆过滤器作为一种集合查询的数据结构,在达到其高效简洁表示集合的同时, 却存在某元素不属于数据集合而被指称属于该数据集合的可能性,即假阳性误判,而不存在假阴性误判(属于集合中的元素而误判为不属于集合中)。目前布鲁姆过滤器查询算法主要有:如标准布鲁姆过滤器算法,计数器布鲁姆过滤器算法(Counting bloom filter(CBF))将位数组中的每一位扩展为一个 counter,从而支持了元素的删除操作。文献提出了一种基于标准布鲁姆过滤器的缓存设计, 实现了可以从缓存收回废弃空间的能力。但是此结构引起假阳性误判率高于标准布鲁姆过滤器。 文献中,Haoyu song 等设想了一种布鲁姆过滤器的扩展型,描述了一种新颖的哈希表结构和查找算法,可以将普通布鲁姆过滤器转化为一种关联哈希桶的计数器布鲁姆过滤器, 通过减少查找过程对内存操作的次数而大幅度提高数据查找效率。

在 EBF 中,每个元素的插入都需要 k 个链表节点用于保存该元素信息,因而占用过多内存过大。而 SFHT 解决了 EBF 中用于保存元素的链表节点的冗余问题, 对于每个元素只需一个链表节点来存储即可。然而由于元素本身的往往占用好几个字节甚至几十个字节,因而系统对内存消耗也不小。然而,这两种典型的布鲁姆过滤器都存在一个缺陷, 每个元素本身都要占据一定的内存,而且其占用内存的大小与元素本身密切相关,这样随着元素的增加使得元素集合逐渐增大, 内存的过大消耗将会使系统难以承受。

本文针对以上的扩展式布鲁姆过滤器结构存在的元素对内存占用的问题, 提出了一种基于值域哈希二次过滤的布鲁姆过滤器(VHBF)。实验表明,VHBF 能获得降低的假阳性误判率。

2 值域哈希布鲁姆过滤器

针对上述布鲁姆过滤器存在的链表节点过大消耗内存的问题, 我们设计了一种基于值域哈希的布鲁姆过滤器特征匹配

张生华: 硕士

基金项目: 国家自然科学基金项目; 基金申请人: 秦 拯,张大方; 项目名称: "基于端系统的网络在线测量理论与方法研究"(No.60273070); 广东省科技项目; 基金申请人: 秦 拯(No.0711020400157,No.7007730); 东莞市科技项目; 基金申请人: 秦 拯 张忠志(No.2007108101021,No.2006101101032)

引擎,如图 1 所示。

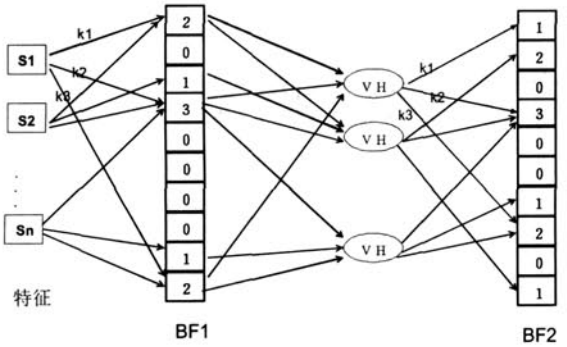

图 1 值域哈希布鲁姆过滤器原理

2.1 VHBF 原理

设特征集合 S={ S1,S2,…Sn }共有 n 个特征,通过 k 个散列函数 kl,k2,…,kk 映射到长度为 m 的向量 BF1 中。每一个散列函数相互独立且函数的取值范围为{0,1,2,…,m-1}。

特征集合映射到 VHBF 时,首先需要将向量 BF1 和 BF2 初始化,向量 BF1 和 BF2 所有位置为 0。注意,BF1 和 BF2 中的每一个位置也称为一个桶(bucket),当元素插入集合中时,对于每一个元素 Si, 需要经过 k 个相互独立的哈希函数计算,Hash_j(si)(1≤j≤k),对于(1≤j≤k),Hash_j(si)=vj,则令 BF1[vj]的值加 1,这样,si 经过 k 个散列函数运算后得到的 k 个结果 vj(1≤j≤k),BF1 中的 k 个位置都自增 1, 此 k 个位置结果即为 VHBF 的中间信息。然后将这 k 个结果进行一次链接操作,即合并为一个 Si 的特征映像 Ti,再对此特征映像 Ti 进行 k 次哈希, V_Hash(vj)(1≤j≤k), 即图 1 中的 VH 运算,k 个结果将使对应于向量 BF2 中的 k 个值自增 1,这样当所有特征重复此过程后,就完成了特征集合的初始化。对于特征集合初始化过程中的每一个特征的初始化,则称为特征的插入。删除特征的操作与插入相似,只是将哈希结果的相应位置的值进行每次减 1 操作即可。

当查询特征是否在集合中时,对于给定的特征 x, 检查向量 BF1 的 k 个位置 Hash_j(si)=vj,(1≤j≤k)是否都为 1。如果有一个为 0,则 x 一定不在集合中;如果全都大于 0, 证明特征可能在集合中,此时将查询得到的中间信息 Hash_j(si)=vj,(1≤j≤k)进行值域哈希过滤运算,即 V_Hash(Hash_j(x)),(1≤j≤k),然后直接在向量 BF2 中的相应位置对比其结果,如果 k 个值都大于 0,则说明特征 x 以一定的假阳性误判率(False Positive)存在特征集合中。下面对此假阳性误判率进行分析。

2.2 VHBF 假阳性误判率分析

假定有 n 个特征,对于 m 个哈希桶中的每个桶(bucket),即每一个位置,其被一个特征置位的概率为 P_set 那么

$$P\_set= 1-(1-\frac{1}{m})^k \quad (1)$$

某个哈希桶被 i 个特征置位的概率为 Pi, 那么设 k 为哈希函数个数,n 个特征全部完成插入操作后

$$Pi= C_n^i(1-(1-\frac{1}{m})^k)^i(1-\frac{1}{m})^{k(n-i)} \quad (2)$$

那么平均哈希桶的长度即平均 counter 值为

$$C_{avg}=\sum_{i=0}^{n} i\times P_i = n(1-(1-\frac{1}{m})^k) \quad (3)$$

对于布鲁姆过滤器,为获得最低的误判率,有下式:

$$k=(m\ln 2/n) \quad (4)$$

$$C_{avg}=\sum_{i=0}^{n} i\times P_i = n(1-(1-\frac{1}{m})^k) = n(1-(1-\frac{1}{m})^{(m\ln 2/n)}) \approx \ln 2 \quad (5)$$

对于 BF1 的假阳性误判率为

$$P_1=(1-(1-\frac{1}{m})^{kn})^k \approx (1-e^{-kn/m})^k \ (m,n\to\infty) \quad (6)$$

BF2 的假阳性误判率为

$$P_2=(1-(1-\frac{1}{S})^m)^r \approx P_2=(1-e^{-m/s})^r (s,n\to\infty) \quad (7)$$

由于 VHBF 的假阳性误判率需要 BF1 和 BF2 共同决定因此,

$$P=(1-(1-\frac{1}{m})^{kn})^k\times(1-(1-\frac{1}{S})^m)^r \quad (8)$$

当 S=m,r=k 时,有 $P=(P_1)^2$。可见 VHBF 的假阳性误判率远低于 EBF 的假阳性误判率。

3 实验分析

为了测试 VHBF 的结构与算法的有效性, 我们对该算法进行了系统仿真实验。在 Linux c 语言编译环境下,我们选取了不同的字符串作为特征集合,我们对 VHBF 和 EBF 的假阳性误判率也进行了比较,如图 2 所示。

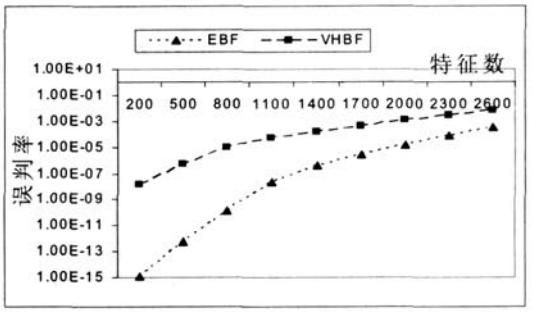

图 2 假阳性误判率分析

取 S=m,r=k,可见 VHBF 的误判率远低于 EBF 的误判率,由 $P=(P1)^2$ 可知 VHBF 的误判率是 EBF 的误判率的平方,因此其误判率的曲线下降更迅速。EBF 和 PFHT 由于这两种结构都需要不同程度的保存原始的特征元素。因此一旦特征元素的数量大规模扩展时,对内存的消耗将不可忽视。我们设计的 VHBF 虽然在准确性上稍逊这两种结构, 但是 VHBF 不需要保存原始的特征元素,可以进行大规模的特征数量扩展,对内存的消耗很低,而且就准确性来说,VHBF 的误判率降到了 EBF 的误判率的平方。

在对内存的存储效率方面, 由于 EBF 与 PFHT 都需要保存元素本身, 而 VHBF 只需要保存其经过二次哈希过滤后的值即可,因而大大地减少了内存占用,而且它的内存占用与特征本身的大小无关。

4 结束语

本文在研究国内外现有布鲁姆过滤器算法的基础上, 针对布鲁姆过滤器的假阳性误判率高和因改善假阳性误判率而引起的元素内存占用问题, 提出了一种基于值域哈希的二次过滤算法, 解决了扩展型布鲁姆过滤器和共享节点快速哈希表的节点对内存的占用问题,保持了布鲁姆过滤器简洁、高效、内存占用低。

本文创新点:提出了一种基于值域哈希的二次过滤方法,提高了布鲁姆过滤器在进行特征检测过程中的内存利用率。

参考文献

[1]甘勇,张勇.网络测量中数据采集系统的设计与实现[J].微计算机信息,2006,12-1:95-97.

[2]李荣鑫.基于智能代理的分布式入侵检测系统模型[J].微计算机信息,2008,2-3:72-73.

(下转第 5 页)

技术创新

VIN 条码,可以查看该批次成车的所有信息,并根据成车状态信息与实际成车状态进行核对;如果有误,则将成车记录标记"返修"状态并将成车转入返修区,如果无误,则打印合格证、保险卡等,并将该成车入库。成车出库模块的主要功能包括:通过扫描成车 VIN 条码和发动机条码,显示的日计划和发运计划对库房现有成车进行校对并录入必要的数据如 "商家代码"、"运输车号"等;校对无误后点击"装箱完毕保存"后系统自动生成出库清单,并对库存进行相应的减库处理,如有误,暂停出库操作并返回入库阶段。角色认证模块的主要功能是用户根据其登录角色访问相应的模块,实现系统基于角色的权限管理。

4 开发成果

图 4 成车条码上线处理

5 结论

成车条码系统投入运行以来,从车间采集的信息更准确、更及时,通过接口将数据定时倒入 ERP 系统后,为 ERP 系统提供了准确的生产线生产状态信息。由于避免了手工键盘方式的信息录入,工人的操作变得简单,查询产品状态信息更加方便,出错率基本杜绝。生产线物流速度明显加快, 日计划完成率百分之百,提高了生产线的计划执行力。上线的成车信息在整个生产线上实现了共享,经过多个环节的校对,保证了成车下线的合格率百分之百。

本文作者创新点: 将条码技术应用在成车装配车间中,并设计和开发成车条码系统,提高成车装配线的效率。此外,实现了车间层数据与 ERP 系统的集成。

参考文献

[1]李国江等.基于 QR Code 条码的飞机加油统计系统[J].微计算机信息,2008,6(2):73-74.
[2]陈蔚芳等.基于条码技术的在制品生产过程管理[J].中国机械工程, 2006.7,17(13):1384-1387.
[3]蒋静等.网络技术与条码技术在电厂物资管理中的应用[J].山西电力[J],2008.6,(3):44-47.
[4]唐万林等.基于 VIN 条码生产过程信息管理的研究与实践[J].机械与电子,2000,(3):18-20.
[5]张勇. 条码识别技术在大型生产企业的应用[J]. 物流技术与应用, 2000,(4):9-11.
[6]胡国仁等.基于条码与 RF 技术的图书仓储管理系统设计[J].天津工业大学学报, 2008.2,27(1):81-83.
[7]王虎等.DataMatrix 二维条码在票务系统中的应用与研究[J].计算机与数字工程,2008,(3):154-156.
[8]姜美莲等. 条码技术在供应链管理中的应用[J]. 轻工机械 2008.4,26(2):116-118.
[9]中国物品编码中心.条码技术与应用[M].北京:清华大学出版社,2003.07:5-8.

作者简介:杨显刚(1978-),男,重庆大学机械工程学院,博士研究生 研究方向:产品数字化设计与制造; 何玉林(1945-),男,重庆大学机械工程学院,本科,博士生导师,主要研究方向为风力发电机组设计技术研究,计算机图形学,产品数字化设计与制造。

Biography:YANG Xian-gang (1978-),Male (Han nationality), Municipality of Chongqing, Mechanical Engneering College of ChongQing University, Ph.D. candidates, Mechanical Engineering ,Product Digital Design and manufacturing.

(400044 重庆市 重庆大学机械学院) 杨显刚 何玉林 刘道双 彭 真

(Chongqing University, Chongqing, 400044, China)
YANG Xian-gang HE Yu-lin LIU Dao-shuang PENG Zhen
通讯地址:(400044 重庆大学机械学院机械基础系) 杨显刚

(收稿日期:2009.03.23)(修稿日期:2009.06.23)

(上接第 92 页)

[3]D. Ficara, S. Giordano, and F. Vitucci. MultiLayer Compressed Counting Bloom Filters. in: proc. of the Conference on Computer Communications[C], INFOCOM,2008 :311 - 315.
[4]Yang Chen, A. Kumar, and Jun Xu, "A New Design of Bloom Filter for Packet Inspection Speedup", Global Telecommunications Conference,2007:1 - 5.
[5]H. Song, J. Lockwood. Multi-pattern signature matching for hardware network intrusion detection systems,IEEE Volume 3, GLOBECOM. 2005:5-9.
[6]H. Song, J. Turner, S. Dharmapurikar, J, Lockwood. Fast Hash Table Lookup Using Extended Bloom Filter: An Aid to Network Processing. In: Proc. of Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications, 2005: 181-192.

作者简介:张生华(1984.1-),男,汉,硕士,湖南大学软件学院。专业:软件工程,研究方向:网络与信息安全; 秦 拯 (1969-),男,汉,博士,湖南大学软件学院教授。研究领域:软件工程,网络安全,可信网络等。

Biography:ZHANG Sheng-hua (1984.1-), male, Master, Software School of Hunan University. Major, Software Engineering; Research area, network and information security.

(410082 长沙 湖南大学软件学院) 张生华 秦 拯
(414000 岳阳 湖南民族职业学院) 宋 勇
(523808 东莞 东莞理工学院) 张忠志

(Software School, Hunan University , Changsha 410082 ,China) ZHANG Sheng-hua QIN Zheng
(Hunan Vocational College for Nationalities, Yueyang Hunan, 414000, China) SONG Yong
(Dongguan University of Technology, Dongguan 523808, China) ZHANG Zhong-zhi
通讯地址:(410082 湖南大学软件学院实验楼 304 室) 张生华

(收稿日期:2009.03.16)(修稿日期:2009.06.16)

技术创新