基于布隆过滤器算法的垃圾邮件地址判别方案

钱曙光，徐 佩，蒲 萌
（长安大学信息工程学院，陕西 西安 710061）

摘要：介绍了布隆过滤器算法在垃圾邮件地址判别方案中的应用，着重描述了布隆过滤器算法的原理以及算法的误判率及解决方案，最后介绍了在垃圾邮件地址判别方案中如何应用布隆过滤器算法。

关键词：布隆过滤器；垃圾邮件；白名单
中图分类号：TP393.08 **文献标识码**：A **文章编号**：1673-1131（2013）03-0029-01

目前网络上充斥着各种各样的垃圾邮件，广告邮件，对于垃圾邮件的地址过滤尤为重要。鉴于垃圾邮件的数量庞大，采用的过滤算法必须能够满足时间和空间上的需求。那么本文采用了 1970 年布隆提出的布隆过滤器算法，这个算法实际上是由一个位数组和一系列随机映射函数组成，用于判别某一个元素是否存在一个集合中。正常情况下，跟哈希判别算法相比，布隆过滤器所需要的空间复杂度只是哈希表的 1/8 或者 1/4，很大程度上节约了计算机内存。

1 算法描述

布隆过滤器是由一个长度为 m 的位数组和 k 个随机映射函数组成，初始状态下将位数组的每一位都设置为 0，如图 1 所示：

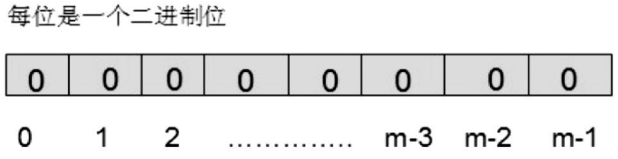

图 1 初始状态大小为 m 的位数组

对于 n 个元素的集合 S={s1，s2……sn}，通过 k 个随机映射函数{f1，f2……fk}，将集合 S 中的每个元素 sj（1≤j≤k）映射为 k 个值{g1，g2……gk}，然后再将位数组 array 中的 array[g1]，array[g2]……array[gk]设置为 1。如下图所示：

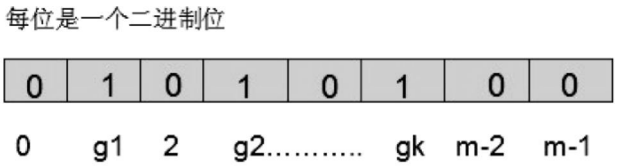

图 2 映射之后的位数组

如果要查找某个元素是否在集合 S 中，则通过映射函数{f1，f2……fk}得到 k 个值{g1，g2……gk}，然后再判断 array[g1]，array[g2]……array[gk]是否都为 1，若全为 1，则元素在集合 S 中，否则不在集合 S 中。这个就是布隆过滤器算法的实现原理。

当然即使 array[g1]，array[g2]……array[gk]都为 1，也不能百分之百代表元素一定在集合 S 中。因为有这个可能 就是集合中的若干个元素通过映射之后得到的数值恰巧包括 g1，g2……gk，那么这种情况下可能会造成误判，但是误判的概率很小。很显然，误判概率的大小跟这 k 个随机函数的映射有关，所以在布隆过滤器中采用随机映射函数的相关性越小越好。还有对于误判的解决方案就是，毕竟被误判的元素数量很微小，那么就可以通过对于经常误判的少量元素设置一个白名单来进行过滤。

2 在垃圾邮件地址判别方案中如何应用布隆过滤器

首先要确定垃圾邮件地址的集合小大 n 和所期望的误判率 p，根据这两个参数能够计算出随机映射函数的个数 k 和位数组的大小 m，他们之间的关系如下：

$$P = 2^{-\ln 2 \frac{m}{n}} \Rightarrow \ln p = \ln 2 \cdot (-\ln 2) \cdot \frac{m}{n} \Rightarrow m = -\frac{n \cdot \ln p}{(\ln 2)^2}$$

$$k = \ln 2 \cdot \frac{m}{n} = 0.7 \cdot \frac{m}{n}$$

可以验证如果 p=0.1，(m/n)=9.6，即存储每个元素需要 9.6bit 位，则 k=0.7*(m/n)=6.72，即存储每个元素需要 9.6 个 bit 位，其中有 6.72 个 bit 位被置为 1 了，因此需要 7 个映射函数。从这里可以看出布隆过滤器的优越性了，存储一个邮件地址，只需要 10 个 bit 位，而用 hash 表存储需要 8*8=64 个 bit 位。

一般情况下，p 和 n 由用户设定，然后根据 p 和 n 的值计算位数组的大小和所需的映射函数的个数，再根据实际情况来设计映射函数。尤其要注意的是，布隆过滤器是不允许删除元素的，因为若删除一个元素，可能会发生漏判的情况。

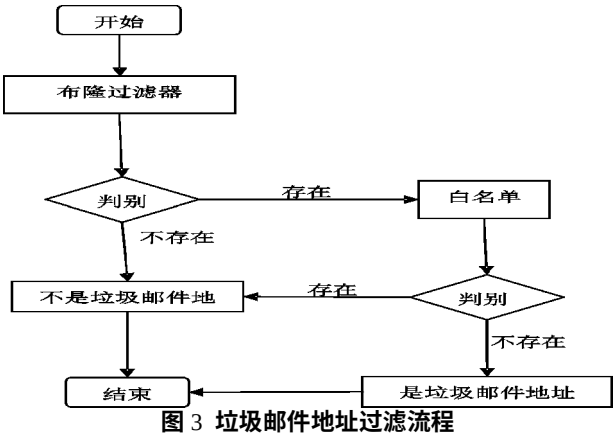

图 3 垃圾邮件地址过滤流程

最后就是对于误判的元素解决方案，因为被误判的元素可以根据元素集合的数量 n 和误判率 p 来计算得到被误判的元素个数即 n*p。由于被误判的元素数量很小，那么对于这少部分被误判的元素可以设置一个白名单，将被误判的元素存放到白名单中，首先对于一个垃圾邮件地址用布隆过滤器进行判别，如果该邮件地址不存在垃圾邮件地址集合中，则可判定为不是垃圾邮件，否则也不能直接判定为垃圾邮件地址，因为有误判率的原因，那么便进行第二次判别，即在白名单中进行查找，如果白名单中不存在该邮件地址，则可以确定该邮件地址是垃圾邮件地址，否则不是。具体判别流程图 3。

参考文献：

[1] 刘威，郭渊博，黄鹏.基于多维布隆过滤器的模式匹配引擎[D].郑州：信息工程大学，2011

作者简介：钱曙光（1989-）男，江苏盐城人，研究生，研究方向为计算机软件与理论；徐佩（1989-），女，湖北武汉人，研究生，研究方向为软件工程；蒲萌（1988-）女，陕西榆林人，研究生，研究方向为交通运输工程。