学校编码：10384　　分类号＿＿密级＿＿
学　　号：20051302298　　UDC＿＿

厦门大学

硕 士 学 位 论 文

基于网络处理器的垃圾邮件过滤系统

Spam Mail Filtering System Based-on Network Processor

林　炼

指导教师姓名：黎忠文　　教授
专 业 名 称：计算机体系结构
论文提交日期：2008　年　4 月
论文答辩时间：2008　年　　月
学位授予日期：2008　年　　月

答辩委员会主席：＿＿＿＿＿＿
评　　阅　　人：＿＿＿＿＿＿

2008 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　月　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1.保密（　），在　　年解密后适用本授权书。

2.不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

摘要

随着互联网的普及，传统的纸质信件已经逐渐被高效、低成本的电子邮件所取代。然而，随之而来的垃圾邮件问题也日趋严重。基于内容的垃圾邮件过滤方法是国内外研究的热点，支持向量机、贝叶斯、Windows 和 KNN 等是其中较为出色的方法，它们有各自优缺点。基于内容的过滤方法在训练邮件样本以及过滤邮件时往往需要耗费大量的时间。

网络处理器是随着网络带宽不断增大，对数据处理的速度要求越来越高的环境下诞生的。拥有多个处理器，针对网络数据处理的指令集以及多级存储结构等特点，使得网络处理器能够高效、便捷的处理 2-7 层的网络数据。

本文在网路处理器上实现了支持向量机和贝叶斯这两种基于内容的垃圾邮件过滤系统。主要工作如下：

1. 介绍了电子邮件相关的协议和标准以及反垃圾邮件技术的现状；分析了支持向量机以及贝叶斯理论在垃圾邮件过滤中的特点。

2. 在 IXP2400 上设计并实现了基于支持向量机和贝叶斯的双分类器垃圾邮件过滤系统。在保证过滤邮件的准确性前提下，利用网络处理器强大的并行处理能力，提升过滤垃圾邮件的速度。

3. 搭建实验平台，并针对 Ling-Spam 语料库对系统进行了性能测试，并对实验结果进行分析。

关键词：垃圾邮件；支持向量机；朴素贝叶斯；网络处理器

Abstract

As the popularization of Internet, the efficient and cheep e-mail has become the substitute to the conventional papery mail. But at the same time, spam mails cause lots of problems. Content-based spam filtering technologies have become the mainstream anti-spam mail methods so far. Support vector machine (SVM), Bayes, windows and KNN are excellent ones of these technologies and they have advantages and disadvantages respectively. The content-based spam filtering methods consume lots of time on training sample mails and classifying testing mails.

As the network bandwidth growing up, higher network data processing rate is required. Network processor is developed to satisfy the requirement. Multi-processors, special instruction set for network data processing and multi-hierarchies memory system make network processor be able to process network data from layer 2 to 7 efficiently and conveniently.

This paper implements content-based spam mail filtering systems based- on SVM and Naïve Bayes on network processor. The major work is as follows:

1. Introducing protocols and standards relative to e-mails as while as the current anti-spam technologies; analyzing the characteristics of SVM and Bayes in spam mail fitering.

2. Implementing a double classifier spam mail fitering system based-on SVM and Bayes. While maintaining the filtering accuracy of content-based method, the system takes advantage of the parallel processing ability of network processor to improve the filtering speed.

3. Constructing the experiment environment and testing the performance of the system on Ling-Spam mail library; analyzing the results of the tests.

Key Words: spam; SVM; Naïve Bayes; Network Processor

目录

Contents