

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

厦门大学博硕士学位论文摘要库

厦门大学学位论文原创性声明

兹提交的学位论文，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

厦门大学博硕士学位论文摘要库

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

厦门大学博硕士学位论文摘要库

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。

本文在网络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

厦门大学博硕士学位论文摘要库

Abstract

As the popularization of Internet, the efficient and cheap 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 filtering.
2. Implementing a double classifier spam mail filtering 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

厦门大学博硕士学位论文摘要库

目录

摘要	1
Abstract	iii
目录	v
Contents	ix
第一章 绪论	1
1.1 垃圾邮件的研究背景	1
1.2 本文的主要工作及意义	3
1.3 本文的内容安排	4
第二章 垃圾邮件过滤技术概述	5
2.1 电子邮件相关协议和标准	5
2.2 电子邮件过滤技术概述	8
2.2.1 基于关键字的过滤	8
2.2.2 基于黑白名单的过滤	9
2.2.3 基于规则的过滤	9
2.2.4 基于内容的过滤的过滤	10
2.3 电子邮件的表示	11
第三章 SVM 和 Bayes 理论	15
3.1 SVM 理论	15
3.2 Bayes 理论	20
第四章 网络处理器	25
4.1 网络处理器 (Network Processor, NP) 概述	25
4.2 Intel 网络处理器 IXP2400 的硬件结构	27
4.3 Intel 网络处理器的 IXA 软件开发框架	31
4.4 网络处理器上程序设计模型	33

第五章 基于网络处理器的垃圾邮件过滤系统的设计与实现	37
5.1 系统模型	37
5.2 offline 模块	38
5.2.1 特征项选择模块	38
5.2.2 构造状态转移表模块.....	40
5.2.3 邮件表示模块	41
5.2.3 Scaling 模块	42
5.2.4 训练模块	44
5.3 online 模块	44
5.3.1 online 子模块设计原则.....	44
5.3.2 接收模块和发送模块.....	48
5.3.3 预处理模块	48
5.3.4 Scaling 模块	52
5.3.5 预测模块	53
5.4 online 与 offline 模块的结合	55
第六章 系统性能测试与分析	57
6.1 网络处理器的安装配置	57
6.2 系统测试环境	60
6.3 判别系统的性能测试与分析	62
6.3.1 SVM 单独判别的性能.....	62
6.3.2 Naïve Bayes 单独判别的性能.....	62
6.3.3 两者共同的准确率.....	63
6.4 多引擎多线程性能测试与分析	63
6.4.1 预处理模块单、多线程性能比较	63
6.4.2 scaling 模块单、多线程性能比较	65
6.4.3 预测模块单、多引擎性能比较	66
6.4.4 系统整体性能分析.....	67
第七章 结束语	69
参考文献	71

附录	73
攻读硕士学位期间的研究成果	74
致谢	75

厦门大学博硕士论文摘要库

厦门大学博硕士学位论文摘要库

Contents

Chinese Abstract	1
English Abstract	iii
Chinese Contents	v
English Contents	ix
Chapter1 Introduction	1
1.1 Background of Spam Mail Filtering	1
1.2 Main Work and Meaning	3
1.3 Content Arrangement	4
Chapter2 Overview of Spam Mail Filtering	5
2.1 Protocols and Standards of E-mail	5
2.2 Overview of Spam Mail Filtering Technologies	8
2.2.1 Key Words-Based Filtering	8
2.2.2 Black, White List-Based Filtering	9
2.2.3 Rule-Based Filtering	9
2.2.4 Content-Based Filtering	10
2.3 E-mail Denoting	11
Chapter3 SVM and Bayes theories	15
3.1 SVM	15
3.2 Bayes	20
Chapter4 Network Processor	25
4.1 Overview	25
4.2 Hardware Architecture of Intel IXP2400	27
4.3 Intel IXA Programming Frame	31
4.4 Programming on IXP2400	33

Chapter5 Designing and Implement of Spam Mail Filtering System on Network Processor	37
5.1 System Model	37
5.2 Offline Module	38
5.2.1 Feature Selecting.....	38
5.2.2 State Transferring Table Constructing	40
5.2.3 E-mail Denoting	41
5.2.3 Scaling	42
5.2.4 Training	44
5.3 Online Module	44
5.3.1 Designing Principle.....	44
5.3.2 Reveiving and Transmitting	48
5.3.3 Preprocessing.....	48
5.3.4 Scaling	52
5.3.5 Predicting	53
5.4 Combination of Offline and Online Modules	55
Chapter6 Performance	57
6.1 Installation and Configuration of Network Processor	57
6.2 Testing Environment	60
6.3 Performance	62
6.3.1 SVM Predicting	62
6.3.2 Naïve Bayes Predicting	62
6.3.3 SVM and Naïve Bayes Predicting	63
6.4 Multi-Micro Engine and Multi-Thread Performance	63
6.4.1 Sigle and Multi Thread Preprocessing Compare	63
6.4.2 Sigle and Multi Thread Scaling Compare.....	65
6.4.3 Sigle and Multi Micro Engine Predicting Compare	66
6.4.4 System Performance	67
Chapter7 Conclusion and Future Work	69

Degree papers are in the "[Xiamen University Electronic Theses and Dissertations Database](#)". Full texts are available in the following ways:

1. If your library is a CALIS member libraries, please log on <http://etd.calis.edu.cn/> and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库