基于 K-近邻法及移动 agent 技术的垃圾邮件检测系统研究 *

王 龙[1]，李晓光[1]，钟绍春[2]
（1. 辽宁大学 信息学院，沈阳 110036；2. 东北师范大学 理想信息技术研究院，长春 130024）

摘 要：为了解决日益严重的垃圾邮件问题，设计了一个新型的基于 K-近邻法及移动 agent 技术的垃圾邮件检测系统。简单介绍了 K-近邻法及移动 agent 技术，详细阐述了基于 K-近邻法及移动 agent 技术的垃圾邮件检测系统的体系结构、工作流程和关键技术。实验结果表明，与同类系统相比，该系统执行速度提高了，对网络稳定性的要求比较低，能够有效阻止垃圾邮件的传播。

关键词：K-近邻法；移动代理；垃圾邮件；垃圾邮件检测
中图分类号：TP393 **文献标志码**：A **文章编号**：1001-3695(2009)07-2630-03
doi:10.3969/j.issn.1001-3695.2009.07.064

Research of spam detection system based on KNN and mobile agent

WANG Long[1], LI Xiao-guang[1], ZHONG Shao-chun[2]
(1. *College of Information, Liaoning University, Shenyang* 110036, *China*; 2. *Institute of Ideal Information Technology, Northeast Normal University, Changchun* 130024, *China*)

Abstract: For solving the growing problem of spam, designed and implemented a new spam detection system based on mobile agent, introduced the relevant technology and the structure of this system, presented some key technology in the process of implementation were presented. By experimental simulations, the test result validated the purpose of this system for spam detecting.

Key words: K-nearest neighbor(KNN); mobile agent; spam; spam detection

邮件检测主要是将垃圾邮件从邮件中过滤出去，邮件的过滤技术主要包括基于黑白名单的过滤、基于邮件头信息分析的过滤和基于邮件内容的过滤等，邮件系统的服务器端和客户端分别通过上述方法对垃圾邮件进行多重过滤。其中基于邮件内容的过滤是目前反垃圾邮件用到的主要技术，是邮件过滤过程中最重要的一步[1~3]。基于邮件内容的过滤方法[4]主要有基于规则和基于概率的过滤方法。其中 K-近邻法(KNN)[5]就是一种基于概率的过滤方法。由于基于概率的过滤方法相对基于规则的过滤方法具有更好的自主性和运行效率，本文将采用 K-近邻法实现邮件内容的过滤。

随着邮件数量的增多，对邮件系统服务器端的计算能力的要求越来越高，单一服务器已无法达到要求。另外，目前大多数垃圾邮件系统的服务器端和客户端各自独立工作，没有形成一个协调统一的整体。本文旨在利用移动 agent 技术，将系统服务器端的各个功能分散到不同服务器上运行，同时将系统的服务器端与客户端有效地联系起来，组成一个整体，达到信息共享和协同工作的目的。

1 KNN 过滤方法和移动 agent 技术

1.1 KNN 过滤方法[5]

在各种分类方法中，KNN 分类法比较简单，能准确地对文本进行分类，因此在垃圾邮件检测中得到广泛应用。

KNN 垃圾邮件过滤方法的工作原理是：先将已有的邮件分为垃圾邮件集合和合法邮件集合，提取每封邮件的加权特征向量，形成训练集。当接收一封新的邮件时处理邮件内容，建立邮件内容的向量空间，最后根据 KNN 算法分别计算出邮件属于垃圾邮件的权重 P_1 和属于合法邮件的权重 P_2。如果 $P_1 > P_2$，则邮件为垃圾邮件；否则为合法邮件。具体算法为：

假设一封邮件文本的特征项表示为 d，训练集 S 共有垃圾邮件 m_1 和合法邮件 m_2 两类，该邮件与训练集中各邮件的相似度为

$$\mathrm{sim}(d,s_i)=\sum_{j=1}^{n} d.t_j \times s_i.t_j / \sqrt{\left(\sum_{j=1}^{n}(d.t_j)^2\right)\times\left(\sum_{j=1}^{n}(s_i.t_j)^2\right)}$$

其中：n 为特征项个数；s_i 为 S 中的一封邮件文本的特征项表示；$d.t_j$ 为特征 j 在 d 中的权重；$s_i.t_j$ 为特征 j 在 s_i 中的权重。

选取相似度最大的 k 封邮件构成近邻集 K，则该邮件属于垃圾邮件的权重 P_1 和属于合法邮件的权重 P_2 为

$$P_l=\sum_{s_i\in K}\mathrm{sim}(d,s_i)y(s_i,m_l)-b_l;l=1,2$$

其中：$y(s_i,m_l)$ 的取值为 1 或 0，当 s_i 属于 m_l 时取 1，反之取 2；b_l 为预先计算测定的 m_l 的阈值。

最后根据 P_1 和 P_2 的值确定该邮件属于垃圾邮件还是合法邮件。

收稿日期：2008-10-09；**修回日期**：2008-11-24 **基金项目**：国家自然科学基金资助项目(60703068)

作者简介：王龙(1978-)，男，吉林吉林人，博士研究生，主要研究方向为机器学习、数据挖掘(wanglong@nenu.edu.cn)；李晓光(1973-)，男，辽宁沈阳人，副教授，博士，主要研究方向为数据挖掘、信息检索、流数据分析；钟绍春(1965-)，男，吉林双辽人，教授，博导，博士，主要研究方向为不确定推理、数据挖掘、信息化教育.

1.2　移动 agent 技术[6]

移动 agent 是一种代替人或者其他程序执行某种任务的程序,它能够在复杂的网络中从一台主机移动到另一台主机中去,移动时该程序可以根据要求挂起其运行,然后转移到网络中的其他地方重新开始或继续执行,最后返回计算结果。

移动 agent 系统包括两个部分:移动 agent 和移动 agent 服务环境。移动 agent 服务环境实现移动 agent 在主机间的迁移,并为其建立远程执行环境。移动 agent 在服务环境中执行并通过 agent 通信语言(agent communication language,ACL)与其他服务器通信或者获得其他 agent 所提供的服务。

移动 agent 的特点之一是移动性,这是它与一般 agent 的区别所在,移动 agent 的移动经常会在异构操作系统的机器之间持续迁移。由于移动 agent 会在运行状态下挂起、迁移,移动的对象除了程序外还必须有其当前的运行状态信息和相应的数据。移动 agent 的另一个特点是其自主性,移动 agent 能够在没有人或其他 agent 的直接干涉和指导下持续地运行,并能够控制其内部状态和动作。Agent 的移动一般是自主决定的。

移动 agent 的这两个特点决定由其构建的应用程序系统具有良好的分布性、智能性和异构性,这正好可以解决目前多数垃圾邮件过滤系统在自主性、信息共享性和协同工作方面存在的不足。因此,将移动 agent 技术应用于垃圾邮件过滤系统完全可行,并且具有非常好的实用价值。

2　系统模型和工作流程

2.1　系统模型

根据垃圾邮件检测系统的功能需求,应用移动 agent 技术及建模方法,分析系统中的组织关系、协作关系及业务关系,将系统中的各类角色和服务设施抽象为相应的移动 agent,设计出系统的模型,如图 1 所示。

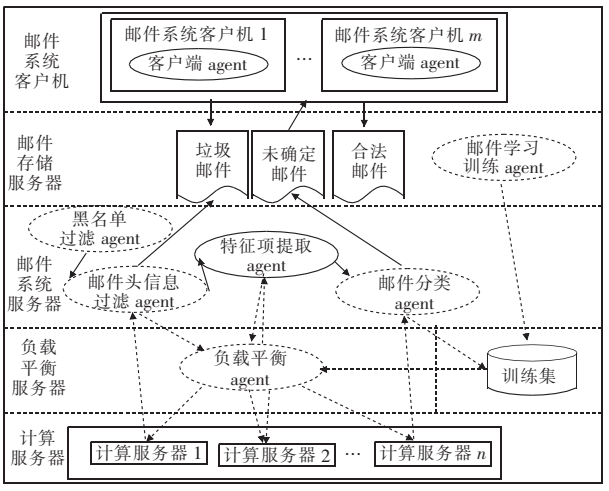

图 1　系统模型

其中邮件头信息过滤 agent、邮件特征项提取 agent、邮件分类 agent、邮件学习训练 agent 都是移动 agent。

2.1.1　邮件系统服务器

邮件系统服务器是一个多 agent 系统,其中邮件头信息过滤 agent、邮件特征项提取 agent、邮件分类 agent 是移动 agent,黑名单过滤 agent 是普通的智能 agent。黑名单过滤 agent 提取邮件的发件人地址,如果在黑名单中马上拒绝连接;邮件头信息过滤 agent 通过对邮件的来源和接收者的域进行分析,判断是不是有效域、是不是完全限定的域名、是不是符合 RFC822 格式等,如果其中一项不满足就拒绝接收;邮件特征项提取 agent对邮件的内容进行摘要提取,并从摘要中提取邮件的特征项;邮件分类 agent 运用 KNN 过滤方法通过邮件的特征项对邮件的内容进行分类,非垃圾邮件标记为 N(normal),垃圾邮件标记为 S(spam)后发送到邮件存储服务器上的未确定邮件集中。

2.1.2　邮件存储服务器与训练集

邮件存储服务器中包含合法邮件集、垃圾邮件集、未确定邮件集和一个邮件学习训练 agent。邮件学习训练 agent 为移动 agent,它通过对垃圾邮件集和合法邮件集的学习训练形成训练集。

2.1.3　负载平衡服务器与计算服务器

由于服务器端的计算量非常大,每封邮件的头信息过滤、特征项提取、内容分类等计算操作需要分开到不同计算服务器上运行。负载平衡服务器主要包括一个负载平衡 agent,负载平衡 agent 为普通的智能 agent。负载平衡 agent 通过对各个计算服务器计算能力的监听将各操作平均分配到每台计算服务器上,充分利用硬件设备,提高系统的运行速度。

2.1.4　邮件系统客户机

邮件系统客户机中包含一个客户端 agent,客户端 agent 为普通的智能 agent。客户端 agent 从邮件存储服务器中的未确定邮件集中接收新邮件,再进行过滤,将非垃圾邮件存到收件箱中,将垃圾邮件存到垃圾箱中,等待用户审核修改,审核修改后将邮件信息发给数据服务器。其中过滤方法包括标记位过滤、黑名单过滤等。

2.2　系统工作流程

图 2 显示了系统的工作流程,图中清楚地表示出移动agent在系统中的移动及其工作方式,以及与其他 agent 的协作关系。

从图 2 中可看出在系统中,邮件的过滤过程只有一小部分是在线操作,从而使系统支持操作的异步性、灵活性及移动性,节省了系统的运行时间、减小了系统的网络流量、降低了系统对网络稳定性的要求。另外,系统以智能 agent 监控分布式计算环境中各节点的负载变化,实现了整个系统的动态负载平衡。

3　实验讨论

系统完成后,在网络带宽正常和随机剧烈变化两种情况下,分别与基于朴素贝叶斯过滤方法及多 agent 技术的垃圾邮件检测系统和只有一台邮件服务器的采用 K-近邻过滤方法的非分布式的垃圾邮件检测系统进行了比较测试。

测试系统采用 Aglets2.1[7,8] 移动 agent 平台,搭建在一个局域网中。其中 1 台服务器作为邮件系统服务器,3 台 PC 机分别作为邮件存储服务器、数据服务器和负载平衡服务器,5 台 PC 机作为计算服务器,10 台 PC 机作为邮件系统的客户机,另外用 2 台 PC 机发送大量的干扰数据,使可用带宽随机剧烈变化,并且能造成网络阻塞。其中服务器的配置为 2 台 P4 3.0e 的 CPU,2 GB 内存;P4 的 PC 机的配置为 P4 1.8a 的 CPU,512 MB 内存。

实验系统搭建后,在互联网上公开专门的测试邮箱,订阅

收集各种邮件,一共收集了 10 000 封邮件集合,其中垃圾邮件 5 785 封,非垃圾邮件 4 215 封。实验过程如下:

a)分别在垃圾邮件集和非垃圾邮件集中随机抽取 70% 的邮件作为学习集,剩下的邮件合在一起随机抽取 50% 分成 5 份与剩下的邮件形成 6 个测试集。

b)分别通过三个系统对每个测试集进行邮件过滤测试,其中邮件数量较多的测试集最后进行测试,并记录每次的邮件数量和错误数量及执行时间。

c)清空数据,重复 a)b)10 次。

d)计算过滤的正确率和执行时间。

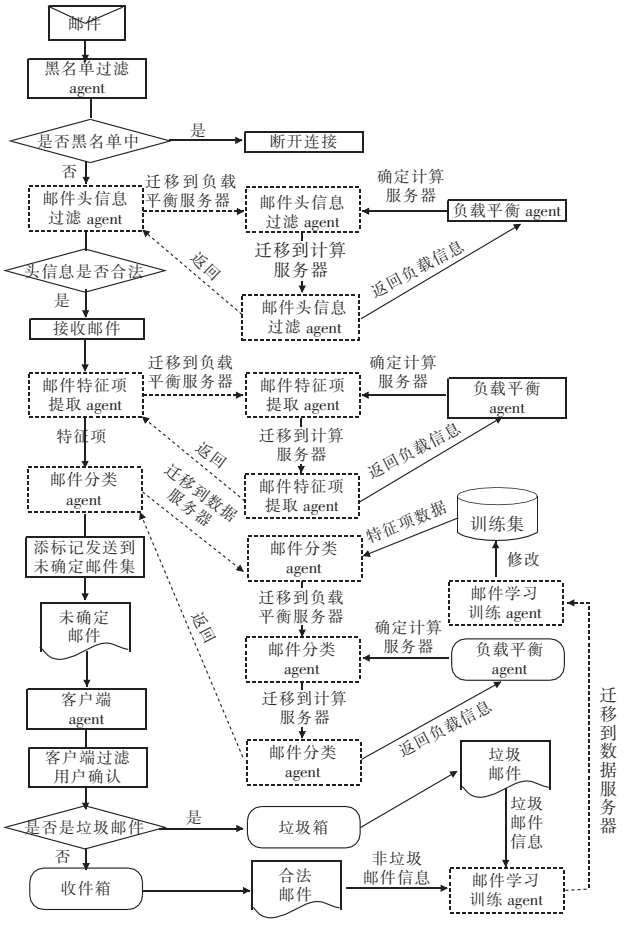

图 2　系统的工作流程

实验结果如图 3 ~ 6 所示。图中“系统 1”为基于 K-近邻法及移动 agent 技术的垃圾邮件检测系统;“系统 2”为基于朴素贝叶斯过滤方法及多 agent 技术的垃圾邮件检测系统;“系统 3”为只有一台邮件服务器的采用 K-近邻过滤方法的非分布式垃圾邮件检测系统。

从以上数据可以得到如下实验结论:

a)本系统的正确率始终保持在 91% 以上,K-近邻过滤方法的正确率与朴素贝叶斯过滤方法的正确率相差无几,基本可以满足用户的需求。

b)测试集 1 ~ 6 的正确率处于上升的趋势。这说明随着系统使用,训练集中的邮件数量不断增加,邮件检测的正确率会不断提高。

c)在网络带宽正常的情况下,本系统与基于朴素贝叶斯过滤方法及多 agent 技术的垃圾邮件检测系统的执行速度要高于只有一台邮件服务器的采用 K-近邻过滤方法的非分布式的垃圾邮件检测系统。特别是对于邮件数量较多的测试集 6,本系统的用时仅为只有一台邮件服务器的采用 K-近邻过滤方法的非分布式垃圾邮件检测系统的 1/3 左右。这说明由于采用了分布式的设计,系统运行速度较快,系统的承受量较大。

d)在网络带宽剧烈变化的情况下,本系统执行速度明显高于基于朴素贝叶斯过滤方法及多 agent 技术的垃圾邮件检测系统和只有一台邮件服务器的采用 K-近邻过滤方法的非分布式的垃圾邮件检测系统。对于邮件数量较多的测试集 6,本系统的用时仅为基于朴素贝叶斯过滤方法及多 agent 技术的垃圾邮件检测系统的 1/7 左右。这说明由于采用了移动 agent 技术,与采用普通的分布式技术相比,网络稳定性对系统的影响较小。

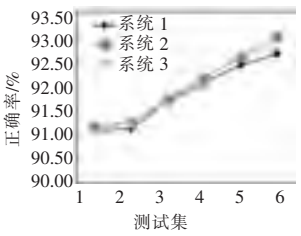

图 3　网络带宽正常情况下的正确率测试

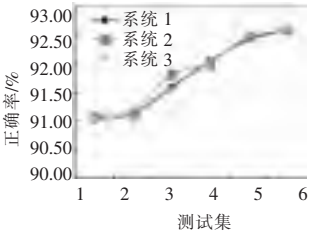

图 4　网络带宽剧烈变化情况下的正确率测试

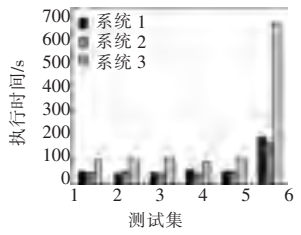

图 5　网络带宽正常情况下的执行速度测试

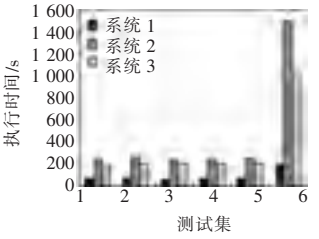

图 6　网络带宽剧烈变化情况下的执行速度测试

4　结束语

本文以 KNN 过滤方法和移动 agent 技术为基础构造了一个分布式的垃圾邮件检测系统,该系统检测准确率高、执行速度快。实验表明,与非分布式的垃圾邮件检测系统相比,该系统具有运行速度快、承受量大的优点;与采用普通的分布式技术的垃圾邮件检测系统相比,该系统对网络稳定性要求比较低。本系统能够有效地控制垃圾邮件的传播,完全适合各种邮件系统用户的需求。

参考文献:

[1] 张铭锋,李云春,李巍. 垃圾邮件过滤的贝叶斯方法综述[J]. 计算机应用研究,2005,22(8):14-19.

[2] 张泽明,罗文坚,王煦法. 一种基于人工免疫的多层垃圾邮件过滤算法[J]. 电子学报,2006,34(9):1616-1620.

[3] 秦志光,罗琴,张凤荔. 一种混合的垃圾邮件过滤算法研究[J]. 电子科技大学学报,2007,36(3):485-488.

[4] 王斌,潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报,2005,19(5):1-10.

[5] 张宁,贾自艳,史忠植. 使用 KNN 算法的文本分类[J]. 计算机工程,2005,31(8):171-172,185.

[6] 张云勇. 移动 agent 及其应用[M]. 北京:清华大学出版社,2002.

[7] WONG D,PACIOREK N,MOORE D. Java-based mobile agents[J]. Communications of the ACM,1999,42(3):92-102.

[8] Aglets software development kit(ASDK)[EB/OL]. http://aglets.trl.ibm.com.