电子文档挖掘技术在电子邮件监控系统中的应用*)

蔡立军　张大方
(湖南大学计算机与通信学院　长沙410082)

The Application of the Electronic Documentary Mining Technology to the E-mail Monitoring System

CAI Li-Jun　ZHANG Da-Fang
(College of Computer and Telecommunication, Hunan University, Changsha 410082)

Abstract　The e-mail monitoring system is a real-time tool to monitor the content of the e-mail. However, the characteristics (the hypersensitive information) in the e-mail extracted systematically by the existing e-mail monitoring system can not sometimes perfectly reflect the practical conditions, meanwhile, the monitoring model established by it is not perfect enough. Therefore, it is easy to have wrong alarm or fail to alarm. According to such conditions, this paper discusses in great details the application of the electronic documentary mining technology in the e-mail monitoring system, and puts forward to adopt the structural model of the electronic documentary mining technology in the e-mail monitoring system.

Keywords　E-mail monitoring, Electronic documentary mining, Hypersensitive information, Simple Bayes, Intensifying method

1　引言

随着Internet技术的发展，各种网络应用服务越来越多。其中，网络中广泛应用的电子邮件(E-mail)正成为一种快捷而廉价的通信手段。然而，电子邮件在给人们带来很多方便的同时，也产生了一系列的新问题：有人利用电子邮件出卖机密情报、散布色情淫秽内容，更有人利用电子邮件散发反动传单、蛊惑人心、散布不利于国家稳定、安全的言论，同时各种垃圾邮件、邮件炸弹、虚假广告也给用户收发E-mail带来了极大的困扰。

电子邮件监控就是公安、国安、边检机关或高精技术企业根据预先设定的主题、关键词对局域网、Internet上的电子邮件及其附件进行监控，如果发现敏感信息，就需要实施报文拦截、实时阻塞、自动发布安全报告给监控管理人员等响应措施。

电子邮件及其附件实际上都是Internet上的一种电子文档。由于传统的知识发现(KDD, Knowledge Discovery in Database)和数据挖掘(DM, Data Mining)主要是针对关系型数据库[1]，对于非关系型数据库(如电子文档)，大部分数据挖掘系统还无能为力。尽管分布式数据库系统已经得到了广泛研究，但Internet的异构性、开放性和动态性要求人们重新研究数据挖掘的新技术、新方法[2~4]，因此，笔者将数据挖掘的思想引入到Internet上的电子文档信息发现领域，设计了一个电子文档挖掘系统[5]：采取建立Internet服务器文档资料镜像站点的方法，采用基于传统数据挖掘的逆过程，即先在镜像站点中对电子文档(包含电子邮件)进行挖掘，然后再把有用的电子文档资料建库，从而提高用户对信息处理的能力和处理速度。

现有的电子邮件监控系统在提取电子邮件中的特征(敏感信息)时，由于没有利用电子文档挖掘技术，因此有时所提取的邮件敏感信息不能很好地反映实际的情况，所建立的监控模型也不够完善，容易出现误警或漏警，给国家、企业造成安全隐患[6]。而电子文档挖掘在从数据中提取特征与规则方面具有非常大的优势。针对这种情况，本文将电子文档挖掘技术应用于电子邮件监控中，提出了采用电子文档挖掘技术的电子邮件监控系统的结构模型。

2　电子文档挖掘技术的应用

电子文档挖掘技术虽然具有不同的适用范围，但在电子邮件监控中可以综合利用电子文档挖掘的关键技术：

(1)运用文档结构解析的有关算法解析电子邮件的报文结构；

(2)运用电子文档挖掘算法挖掘电子邮件中的敏感信息，并利用Bayes分类器建立电子邮件训练样本集和电子邮件敏感信息库；

(3)基于(1)和(2)的分析结果，对电子邮件中的敏感信息进行描述；

(4)运用电子文档挖掘的自动归档处理技术，可以把挖掘的结果自动生成安全报告，自动整理、归档；

(5)运用电子文档挖掘的自动发布技术，把挖掘出的敏感邮件自动变为Web页面，并发送给邮件监控管理员。

下面对关键问题的处理过程给出如下的解决方案。

2.1　Internet E-mail报文结构解析

Internet E-mail系统的基本组成如图1所示。“用户代理”(UA: User-Agent)主要用于邮件的生成和对邮件进行各种处理。“报文传输代理”(MTA: Messages Transfer Agent)主要负责邮件的传输。输出队列和邮箱主要起缓冲的作用，这样可以使用户收发邮件与实际的邮件传输区分开。电子邮件系统主要涉及的协议有SMTP协议(RFC 821)[7]、822文本协议(RFC 822)[8]、POP3协议(RFC 1725)[9]、MIME(RFC 1521、

*)湖南省教育厅资助项目(01C012)。蔡立军　副教授，研究方向为计算机网络安全、计算机应用。张大方　教授，博士生导师，主要从事可信系统与网络的研究。

RFC 1522)[10,11]协议。

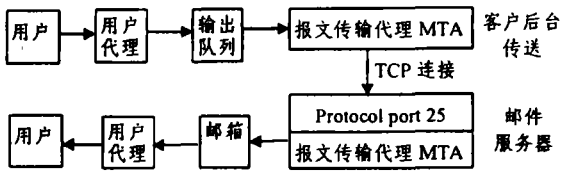

图1 Internet E-mail 系统的基本组成

电子邮件报文结构通常包含三部分:报文传输代理使用的信封(envelope),用户代理使用的头(header),以及包含接收者使用的报文和数据的 E-mail 体(body)。为了解析 E-mail 报文结构,必须要了解 E-mail 报文的生成、发送和接收过程。

(1)E-mail 报文的生成和发送 SMTP 协议规定了客户与服务器 MTAs 之间双向通信的规则,及信封信息的传递。在与 SMTP 服务器(公共端口号25)建立连接的基础上,客户 MTA 把命令发送到服务器 MTA,服务器 MTA 把应答发送回客户 MTA,客户再根据应答进行相应的处理。具体的 E-mail 报文的生成、发送过程如下所示:

S:220 BERKELEY. ARPA Simple Mail Transfer Service Ready //服务器 MTA 准备好;
C:HELO USC-ISIF. ARPA //客户 MTA 发送请求命令;
S:250 BERKELEY. ARPA //服务器 MTA 接受请求;
C:MAIL FROM:〈Postel@USC-ISIF. ARPA〉//客户 MTA 发送反向路径参数;
S:250 OK
C:RCPT TO:〈fabry@BERKELEY. ARPA〉//客户 MTA 发送目的地址;
S:250 OK
C:DATA //客户 MTA 请求发送邮件数据;
S:354 Start mail input;end with〈CRLF〉.〈CRLF〉 //服务器 MTA 接受请求;
C:Blah blah blah… //发送邮件数据;
C:…etc. etc. etc.
C:. //邮件数据发送完毕;
S:250 OK
C:QUIT //退出;
S:221 BERKELEY. ARPA Service closing transmission channel //关闭会话;

其中,MAIL 和 RCPT 命令建立的连接中"FROM:"和"TO:"域的信息便是信封信息,即将数据从一个主机传送给另一个主机的信息。

822文本协议为另外两个 E-mail 报文组成部分,即头和体定义了标准。为解释较复杂的头结构,RFC822使用了如下几个头域:Date,From,To,Subject,Reply-To,Cc,Comment,In-Reply-To,X-Special-Action 和 Message-ID。一般的头格式是一个域名,后跟冒号,接着是文本。如:

Date: 27 Aug 76 0932 PDT
From: Ken Davis〈KDavis@This-Host. This-net〉
Subject: Re:The Syntax in the RFC
Sender: KSecy@Other-Host
To: George Jones〈Group @Some-Reg. An-Org〉, AL. Neuman @MAD. Publisher

Internet E-mail 体典型地使用 NVT ASCII 码。另外用一空行将 E-mail 头和 E-mail 体分开。

在邮件处理期间,用户在用户代理软件中输入 E-mail 数据。接着,用户代理增加头域,然后把体和头传送到 MTA。接下来 MTA 增加信封信息,再把完整的邮件数据包传送给另一个 MTA(或者是目的 MTA,或者是中继代理)。这样便完成了简单的邮件生成与发送。

(2)E-mail 报文的接收 E-mail 报文的接收主要涉及 POP3邮局协议。典型的 POP3会话需经过的三个阶段:即鉴别、处理和更新。POP3客户和服务器(公开端口号 110)建立连接后,会话进入鉴别阶段。在鉴别阶段,客户对服务器标识自己。如果鉴别成功,则服务器就打开客户的邮箱,会话也就进入处理阶段。在处理阶段,客户请求服务器提供信息(如邮件列表)或完成动作(如取走指定的邮件报文)。然后,会话进入更新阶段,在这一阶段结束会话,中断连接。

具体的 E-mail 报文接收过程(略)。

(3)E-mail 报文结构的扩展 由于简单邮件传输协议使用 NVT ASCII 码对数据进行编码,因此不能传送多媒体数据(如图像,声音和视频等)。RFC1452中定义了 ESMTP(扩展的 SMTP),允许用户定义扩展级,任何以字母 X 开始的域,都为用户自定义域。而 MIME 则增加了如下五个新的头域 (MIME-Version; Content-Type; Content-Transfer-Encoding; Content-ID; Content-Description) 到 Internet E-mail 报文中[12],极大地扩展了电子邮件的功能,使其应用范围大大地增加。

2.2 电子文档的挖掘算法

本系统采用双扫描缓冲区的无回溯搜索算法,对于搜索过程采用双栈结构,并且采用两个指示器:扫描指针指向当前识别的字符,匹配指针则指向已经匹配的字符缓冲区。其中扫描缓冲区中存放的是从文件中读取的数据信息,而匹配缓冲区中存放的是要挖掘的关键词。

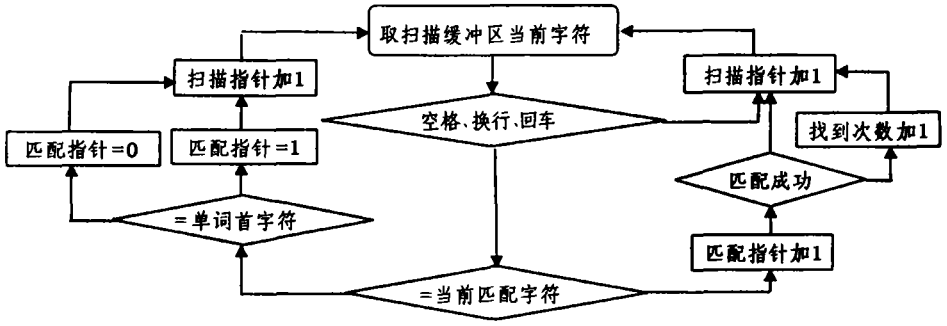

图2 挖掘算法流程图

算法描述如下:定义扫描缓冲区为 Buffer,扫描指针为 nCurPos,关键词缓冲区为 szString,匹配指针为 nHasDone。挖掘算法流程见图2。为了突出挖掘算法,这里没有对从文件中读取数据进行描述,而直接对已经读入到缓冲区中的数据进行处理。

(1)预处理:根据要求,去掉扫描缓冲区中的空格、换行符和回车符。

(2)比较扫描缓冲区和匹配缓冲两个栈中的当前字符,如果相等(区分大小写)则匹配指针加1。当匹配指针指到单词的结束时,增加匹配成功的次数同时将匹配指针复位。

(3)如果两个字符不等,则将匹配指针复位,重新比较。

(4)取出扫描缓冲区中的下一个字符进行比较。

显然,该挖掘算法是判断存储与流的关系,是无回溯的,对所有文件的挖掘只需要打开和读取文件一次,这样就可以大大提高挖掘的速度和精度。

2.3 Bayes 分类器的设计

朴素贝叶斯分类器是一个简单、有效而且在实际使用中很成功的分类器,其性能可以与神经网络、决策树分类器相比,在某些场合优于其他分类器[13]。

设有变量集 $U=\{A_1,\cdots,A_n,C\}$,其中 $A_1,\cdots,A_n$ 是实例的属性变量,C 是取 m 个值的类变量。假设所有的属性条件都独立于类变量 C,即每一个属性变量都以类变量作为唯一的父结点,就得到朴素的贝叶斯分类器。而增强型方法的主要思想是从训练例学习一系列的分类器,每一个分类器根据前一个分类器错误分类的实例、对训练集的权重进行修正,再学习新的分类器。对朴素贝叶斯分类器采用增强型方法,其性能一般说优于已经发表的使用其他学习方法的最好的结果。增强型方法的时间复杂度为 O(Tef),其中 T 是增强的次数,e 是训练例的个数,f 是属性的个数。

电子邮件监控系统使用朴素贝叶斯分类器对邮件进行分类过滤,先设定属性变量个数 n(n 可取在1000左右),每一个属性变量与电子邮件敏感信息库中的每一个特征相对应;类变量 C 有两个,即敏感邮件及非敏感邮件。对于特征表中的每一个特征,若它没有在某封邮件中出现,则其对应的属性变量值 $a_i=0$;反之 $a_i=1$。由于将一封普通邮件归为敏感邮件所带来的危害远小于将一封敏感邮件归为普通邮件,因此阈值 L 不要设得太大,如70%。邮件的后验概率只有在大于这个 L 值时,才认为这是一封敏感邮件。

为了提高朴素贝叶斯分类器的性能,使用了增强型方法。在使用这种方法的时候,必须确定增强的总趟数 T。T 的确定方法是:首先一直对朴素贝叶斯进行增强,直到样本的分类出错率小于一个很小的值,比如10^{-3},记下此时的趟数 T_0;然后再进行 $T_0/10$次增强,即总趟数 $T=1.1T_0$。

3 电子邮件监控系统的结构

在电子邮件监控系统运用电子文档挖掘技术可以有效地从各种数据中提取出有用的信息。所以在电子邮件监控系统中,在建立电子邮件训练样本集和敏感信息库时,以及对当前用户的电子邮件进行监控时,都可以采用电子文档挖掘中相关的技术。

本文根据电子邮件监控系统和电子文档挖掘技术的特征,提出了一种运用电子文档挖掘技术的电子邮件监控系统的结构。其处理过程为:首先在镜像站点中获得用户的历史电子邮件数据,并对电子邮件的报文结构进行解析、分类、挖掘,通过对电子邮件进行样本训练,获得敏感信息的特征模式,再以该模式为基础,采用朴素贝叶斯分类器和增强型方法进行机器学习,最终获得一个电子邮件监控器。引入电子文档挖掘技术的电子邮件监控系统的结构如图3所示。其结构说明如下:

(1)电子文档挖掘:先对电子邮件的报文结构进行解析、分类,接着采用双扫描缓冲区的无回溯搜索算法对电子邮件进行挖掘处理,提取敏感信息,再把挖掘的结果返回给操作者或电子邮件监控器。

(2)敏感信息提取:采用多进程并根据用户任意配置的挖掘主题、关键词等调用不同的挖掘程序分别对不同的电子邮件及其附件进行挖掘处理所得到的一个 d 维向量 W 来表示敏感邮件的样本特征。d 为描述样本特征数,并将其放入敏感信息库中。

(3)朴素贝叶斯分类器:主要是利用先验概率求出后验概率,并且根据训练样本集构造分类器,分类器根据邮件的后验概率对样本进行分类。

(4)增强型方法:用来增强前面获得的朴素贝叶斯分类器的性能,以得到一个最佳的邮件监控器。

(5)敏感信息库:存放系统需要的敏感信息及特定人员的基本信息,如电话号码、家庭住址、IP 地址、MAC 地址、上网方式等。电子邮件监控器将当前用户的电子邮件特征与其进行比较判断,从而可以判断出用户的电子邮件是否是敏感邮件。

(6)电子邮件监控器:系统根据一定的算法,从敏感信息库中提取相关规则的数据,对当前用户的电子邮件特征进行监控检测。根据监控检测的结果,做出相应的行动。如果属于敏感邮件,则系统做出报警、自动生成 Web 上的发布清单,并采取报文拦截、实时阻塞等一系列措施,阻止电子邮件的正常发送。如果属于正常的电子邮件,则系统继续对用户的电子邮件进行监控。

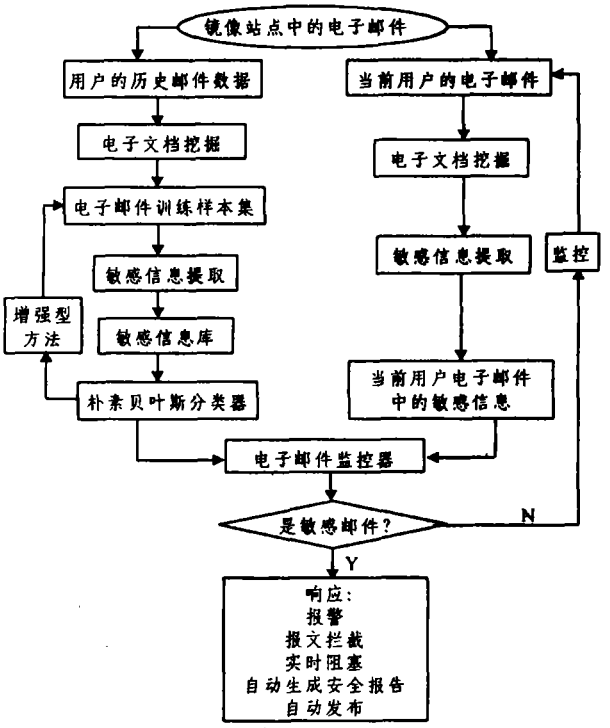

图3 引入电子文档挖掘技术的电子邮件监控系统结构图

结束语 本文根据电子文档挖掘和电子邮件监控的特征,将两者结合在一起,提出了应用电子文档挖掘技术的电子邮件监控系统的结构模型。根据这一结构模型设计的电子邮件监控系统在提取敏感信息和用户邮件特征方面的准确性有了一定的提高,降低了误警或漏警率。

电子邮件监控系统可实际应用于国安、司法、边检、企业等领域,强化公安机关的管理职能,保证公司内部的机密、技术不通过电子邮件外泄,具有良好的应用前景。

参 考 文 献

1 Piatetsky-Shapiro G, Fayyad U, Smity P. From data mining to knowledge discovery: an overview [A]. In: Advances in Knowledge Discovery and Data Mining[C]. Cambridge, Mass: AAA/MIT Press, 1996. 1~34

(下转第78页)

送速度等于 r_i，其缓冲区大小为 Bvi，该路视频流其实就是传统的点播方式。

为达到最佳的磁盘工作效率，缓冲动态调整算法需要考虑网络传输速度和缓存容量。网络传输速度包括单个播送流传输速率和服务器针对所有客户能提供的传输速率，其中单个播送流传输速率由客户端计算能力、客户端数据存取速度、网络带宽及服务端提供的传输速率等因素决定。服务端网络传输速率由允许的网络带宽和服务器数据传输能力决定，它限定了同时播送实时演播视频流的最大传送能力。

4.1 维持已有点播用户

在稳态运行过程当中，视频服务器为 n 个点播用户保持有 k 路视频流，它根据客户端和网络传输情况为 k 路视频流尽力读取并传送视频数据。视频服务器总的视频传送速度为 V，其中第 i 路视频流的传送速度为 v_i。当单路视频流的传送速率大于其数据读取速率时，其从磁盘读出的视频数据立即被传送到客户端。而当该视频流的传送速率小于其数据读取速率时，其对应的动态缓冲区将不断增大，每当其缓冲区装满时就必须为新读取的数据申请一个新的页面，并将其加在页面数据链后端。传送成功的页标记为脏，供以后使用。

动态缓冲区大小的下限为 Bvi，根据推算得出其上限为 $((Bs-\sum_{i=1}^{k} Bvi) * v_i)/V$。当第 i 路视频流达到其最大缓存时，磁盘读取被暂停。而当第 j 路视频流传送完毕时，清除其对应的缓冲区，系统在 k-1路视频流的基础上达到一个新的平衡。当网络带宽足够时，磁盘读取速度接近最大速度。

4.2 接受新的点播请求

当收到一个新的点播请求时，系统原有的平衡被打破，进入一个暂态过程。如果视频服务器接受该第 k+1路点播视频流，则系统为维持 k+1路视频流必须同时满足的充要条件有：

$$(1)v_{k+1}>r_{k+1};(2)\sum_{j=1}^{k+1} r_j>r_s;(3)\sum_{j=1}^{k+1} r_j>d_s;$$

$$(4)\sum_{j=1}^{k+1} Bvj>Bs。$$

在资源充分的情况下，系统接收该请求。若视频服务器中有足够多的空闲页，则马上为新加入者分配一个新的动态缓冲区。而当视频服务器中无空闲页面时，则需要从缓冲区大小超过其 Bvi 的各动态缓冲区中抽取页面，被抽取的对象为 $\max\{\frac{Bi-Bvi\} * V}{v_i}\}$，它反映了被抽取的动态缓冲区超长因子和数据传输速度因子的权衡。当抽取到足够的缓存页面后，视频服务器停止抽取页面并为新视频流创建动态缓冲区。为了维持新加入者后的系统运转，系统达成一个新的平衡条件，所有超过新平衡条件上限的旧缓冲区停止增长。经过一段时间的运行后，视频服务器达到一个新的平衡态。此时，各点播视频流的数据读取速率有所下降，但仍能保证其数据率不小于 r_i，即保证该点播视频流数据的磁盘读取速度不小于该视频流实时播放的数据消耗速度。

由于传输速度和缓存受限，已读入缓存的数据有可能被废弃，从而有些数据需二次（或多次）从磁盘读取，降低了磁盘读取效率，出现这种情况说明瓶颈是网络传输速度而不是磁盘读取速度。

4.3 暂缓新的点播请求

如果系统资源不足以保证新的点播请求加入后的正常播送，则拒绝新的请求，并告知该请求何时能获得满足，让用户选择是取消该点播请求还是预定该点播节目。预定一个节目是指视频服务器为该请求分配和保留一些资源，并保证该用户在预定的时间能够实时播放该请求的视频。预定方式工作时，系统能够根据数据相关性控制预先将部分数据传送给用户，提高了系统的利用率。

结束语 基于可控的小区高速网环境，本文提出了一个采用自适应动态缓存的视频点播机制。在服务分类的带宽保证基础上，本策略利用动态缓存极大地提高了视频服务器磁盘的工作效率，缓解了磁盘存取速度对系统性能的影响。在小区的视频点播及虚拟实验室视频流媒体的分发项目中，该算法实现表现出很好的运行效果。由于采用了客户端全缓冲手段，对于授权、认证及记费等问题需要对系统作进一步的研究。

参 考 文 献

1 Gao L, Towsley D. Supplying instantaneous video-on-demand services using controlled multicast. In: Proc. IEEE Intl. Conf. on Multimedia Computing and Systems, 1999

2 Dan A, Sitaram D, Shahabuddin P. Dynamic patching policies for an on-demand video server. Multimedia Systems, 1996,4(3):112～121

3 Dan A, Sitaram D. Buffer management policy for an on-demand video server. IBM Research Division, T J Watson Research Center: [Tech Rep: IBM Research Report RC 19437]. Armank, New York, 1993

4 Chan S-H G, Tobagi F. Distributed Servers Architecture for Networked Video Services. IEEE/ACM Trans. on Networking, 2001,9(2)

5 Brubeck D W, Rowe L A. Hierarchical storage management in a distributed VOD system. IEEE Multimedia, 1996,3(3): 37～47

6 李方敏，李仁发，叶澄清. 一种核心无状态保存的自适应成比例公平带宽分配机制. 计算机研究与发展，2003,39(2)

（上接第60页）

2 Innom W H. Building the Data Warehouse, 2nd[M], 2000

3 Wu K L, Yu P S, Ballman A. SpeedTracer: A Web usage mining and analysis tool. [J] IBM System Journal,1998,37(1): 89～105

4 Cooley R, Srivastava J. Web Mining: information and pattern discovery on the World Wide Web (A Survey Paper). [C] In: Proc. Of the 9th IEEE Intl. Conf. on Tools with Artficial Intelligence (ICTAI'97), Nov. 1997. http://maya. cs. depaul. edu/ ～mobasher/ papers/ webminer-tai97. ps

5 李立明，蔡立军，等. 电子文档信息自动挖掘技术中的预处理研究. 计算技术与自动化，2002,21(2):92～96

6 Hush Communications Corp. Hush Mail FAQ. http://www. hush mail. com/faq. htm, Nov. 1999

7 Postel J. Simple mail transfer protocol. STD 10, RFC821, USC/ Information Sciences Institute, Aug. 1982

8 Crocker D. Standard for the format of ARPA Internet text message. STD 11, RFC 522, UDEL, Aug. 1982

9 Myers J. Post office protocol — version 3. RFC1725, Dover Beach Consulting, Inc., Nov. 1994

10 Borenstein N, Freed N. MIME (Multipurpose Internet Mall Extentions) part one: mechanisms for specifying and describing the format of Internet message bodies. RFC 1521, Bellcore, Innosoft, Sep. 1993

11 Moore K. MIME(Multipurpose Internet Mail Extentions) part two: message header extensions for non—ASCII text. RFC1522, University of Tennessee, Sep. 1993

12 Ramsdell B. S/MIME Version 3 Message Specification. IETF RFC 2633, Jun. 1999

13 Sahami M, Dumais S, Heckerrnan D. A Baysian Approach to Filtering Junk E-M ail. AAAI 98 Workshop on Text Categorization. July 1998