

CCERT 中文垃圾邮件过滤解决方案

CCERT 中文垃圾邮件过滤规则集 Chinese_rules.cf

2005-4-2 版

(<http://www.ccert.edu.cn/spam/sa/CCERT-Anti-Spam-Solutions.pdf>)

陈光英 (Quang-Anh TRAN)

Email: qa@ccert.edu.cn

CCERT-2005

版权所有：中国教育和科研网紧急响应组 (CCERT) 2004-2005

目录

| | |
|-------------------------------------|----|
| 1. Chinese_rules.cf FAQ | 3 |
| 1.1 什么是 Chinese_rules.cf | 3 |
| 1.2 Chinese_rules.cf 的理论背景 | 3 |
| 1.3 Chinese_rules.cf 的生成和使用框架 | 4 |
| 1.4 Chinese_rules.cf 的匹配速度问题 | 4 |
| 1.5 Chinese_rules.cf 的准确率 | 5 |
| 1.6 Chinese_rules.cf 的用户统计 | 5 |
| 1.7 Chinese_rules.cf 的使用方法 | 6 |
| 2. 在 sendmail 系统过滤中文垃圾邮件 | 7 |
| 2.1 框架 | 7 |
| 2.2 安装 SpamAssassin | 7 |
| 2.3 安装 Mimedefang | 8 |
| 2.4 配置 Sendmail | 8 |
| 2.5 安装 Chinese_rules.cf | 9 |
| 2.6 自动更新 Chinese_rules.cf | 9 |
| 2.7 注意 | 10 |
| 3. 在 qmail 系统过滤中文垃圾邮件 | 10 |
| 3.1 框架 | 10 |
| 3.2 安装和配置 qmail | 11 |
| 3.3 安装和配置 SpamAssassin | 11 |
| 3.4 安装 Chinese_rules.cf | 11 |
| 3.5 qmail 与 SpamAssassin 结合 | 11 |
| 4. 在 Windows 系统过滤中文垃圾邮件 | 12 |
| 4.1 框架 | 12 |
| 4.2 安装 pop3proxy | 13 |
| 4.3 安装 SpamAssassin | 13 |
| 4.4 安装 Chinese_rules.cf | 13 |
| 4.5 配置 | 13 |
| 4.6 自动启动 pop3proxy | 14 |
| 4.7 自动更新 Chinese_rules.cf | 14 |
| 5. Outlook Express 设置步骤 | 14 |
| 5.1 建立一个垃圾邮件文件夹名字为 SPAM | 14 |
| 5.2 添加邮件规则把垃圾邮件移到文件夹 SPAM 中 | 15 |
| 5.3 察看垃圾邮件 | 18 |
| 6. 参考文献 | 20 |

1. Chinese_rules.cf FAQ

1.1 什么是 Chinese_rules.cf

Chinese_rules.cf 是用于业界广泛使用的免费垃圾邮件过滤系统 SpamAssassin [1][2] 的中文垃圾邮件过滤规则集。由于以前没有中文的过滤规则集，SpamAssassin 对中文邮件过滤的准确性不高。CCERT[3]反垃圾邮件研究小组利用 CCERT 所掌握的最新和丰富的样本数据，推出了第一个基于 SpamAssassin 的中文垃圾邮件过滤规则集 Chinese_rules.cf [4]。该规则集每周更新一次，时效性非常好。

Chinese_rules.cf 是在 SpamAssassin 官方网站上发布的第一个中文垃圾邮件过滤规则集，也是用 Google, Yahoo, 百度, MSN 搜索“中文垃圾邮件过滤”时所返回的第一条结果。

1.2 Chinese_rules.cf 的理论背景

Chinese_rules.cf 是邮件内容过滤规则集。目前邮件内容过滤技术可以分为两种方法：基于规则和基于统计的方法。基于规则的方法就是在邮件内容中寻找特定的模式，例如主题包含“免费”。基于统计的就是使用统计方法解决邮件的二元分类问题，其中分类机跟据垃圾邮件和正常邮件的样本训练出来。在垃圾邮件过滤技术中最常用的统计方法就是贝叶斯准则。

基于规则方法的优点是规则可以共享，因此它的推广性很强。一个人写出的规则可以提供给多个人，多个服务器使用。然而它的缺点就是更新速度慢。因为规则一般都是人工编写生成，所以新规则的产生速度跟不上新垃圾邮件出现的速度，换句话说，它的时效性较差。

基于统计的方法的优点就是分类机由程序自动训练出来，只要及时更新样本训练集就可以使分类机更新的速度跟得上垃圾邮件出现的速度，即它的时效性很强。然而该方法的缺点就是分类机不能共享，某个用户用自己的邮件样本集训练出来的分类机对其他用户可能效果不佳，因此该方法的推广性较差。

Chinese_rules.cf 使用基于统计规则的新方法，即它所使用的规则是由统计方法自动生成的。该方法吸取了基于规则和基于统计的优点：因为它是一种基于规则的方法，因此推广性很强，又因为它的规则是由统计方法自动生成的，因此它的时效性也很强。Chinese_rules.cf 和传统方法比较如表 1 所示。

表 1: Chinese_rules.cf 和传统方法比较

| | 推广性 | 时效性 |
|------------------|-----|-----|
| 基于规则 | 好 | 差 |
| 基于统计 | 差 | 好 |
| Chinese_rules.cf | 好 | 好 |

CCERT 反垃圾邮件组自从 1998 年成立以来，每天都处理大量的垃圾邮件投诉，掌握最新和最丰富的样本数据。Chinese_rules.cf 就在此最新和最丰富的样本数据库的基础上，通过

统计方法自动产生的。

1.3 Chinese_rules.cf 的生成和使用框架

Chinese_rules.cf 的生成和使用框架如图 1 所示。首先，利用 CCERT 垃圾邮件处理服务和用户反馈信息来维护一个最新，最全的垃圾/正常邮件样本库，再利用统计方法，根据垃圾/正常邮件样本库自动生成规则集 Chinese_rules.cf。因为样本库是最新的，Chinese_rules.cf 的时效性就非常强。CCERT 将该规则集在 CCERT 主页上发布，作为 CCERT 提供的一种对外服务。各地用户（服务器）通过 CCERT 主页下载 Chinese_rules.cf，这样使 Chinese_rules.cf 的推广性很强。

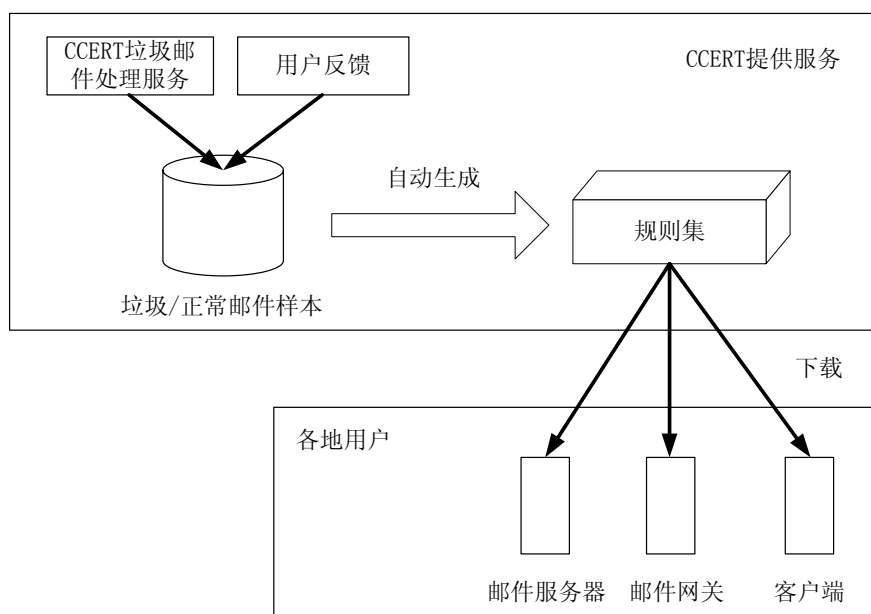


图 1、 Chinese_rules.cf 的生成和使用框架

1.4 Chinese_rules.cf 的匹配速度问题

Chinese_rules.cf 规则集一般被控制在 500 个规则左右。这一数字也许使人对 Chinese_rules.cf 的匹配速度有点置疑。仔细分析和测试结果表明 Chinese_rules.cf 的匹配性能是非常高的，原因是：一、Chinese_rules.cf 的规则都很简单，都是一个比较短的字符串，中间没有带任何一个通配符，这样匹配速度比复杂的规则要快的多；二、Chinese_rules.cf 中有 90%是邮件主题的规则，只有 10%是信体的规则。由于邮件主题往往比较短，因此 Chinese_rules.cf 的匹配速度会比较快。

以上是对性能的理论分析。我们用一台普通 PC(P4 2.8G CPU)，用 Chinese_rules.cf (2004 Dec 21 的版本) 对 178482 封邮件匹配，则结果是平均匹配一封大小为 5.0K 的邮件只需要 0.04 秒。这个结果实非常好的，因为如果一个邮件服务器的邮件平均大小为 5.0K（不算附件），那么只要一台普通 PC 每天就可以处理 216 万封邮件。一般的学生邮件服务器每天收发 30 万封左右。换句话说，只要在现有的邮件服务器上加上如同上述一台 PC 的处理性能就足以满足处理垃圾邮件的硬件需求。

1.5 Chinese_rules.cf 的准确率

Chinese_rules.cf 的每一个版本都带有对准确率的测试结果。以 2005 Jan 2 的版本为例，测试结果如下：

表 2、Chinese_rules.cf 2005 Jan 2 版本的性能

| 阈值 | 垃圾邮件查全率 (共 16729) | 正常邮件误判率 (共 93655) |
|-----|----------------------|----------------------|
| 0.5 | 95.0% | 5.1% |
| 1.0 | 92.9% | 1.6% |
| 1.5 | 90.4% | 0.4% |
| 2.0 | 87.9% | 0.1% |
| 2.5 | 84.5% | 0.0% |
| 3.0 | 81.1% | 0.0% |
| 3.5 | 76.6% | 0.0% |
| 4.0 | 72.4% | 0.0% |
| 4.5 | 67.0% | 0.0% |

表 2 中的结果就是在测试规程中，除了 Chinese_rules.cf 以外不使用其他任何规则。在实际情况，Chinese_rules.cf 一般都会跟 SpamAssassin 的缺省规则同时使用。因为 SpamAssassin 的缺省规则中有一部分是描述邮件行为的规则，对检测中文垃圾邮件起作用，因此实际的性能会比以上实验结果要好。

注意、对于每天处理 40 万封邮件以上的邮件服务器来说，能够容忍的性能就是正常邮件误判率小于 5% 的同时，垃圾邮件的检测率大于 90%。

1.6 Chinese_rules.cf 的用户统计

CCERT 于 2004 年 9 月 7 日在网上发布 Chinese_rules.cf。从 9 月 7 日至 12 月 31 日的用户统计情况如下。图 2 就是用户查看规则集的统计（按 IP）。可以看出规则集的知名度在持续上升。

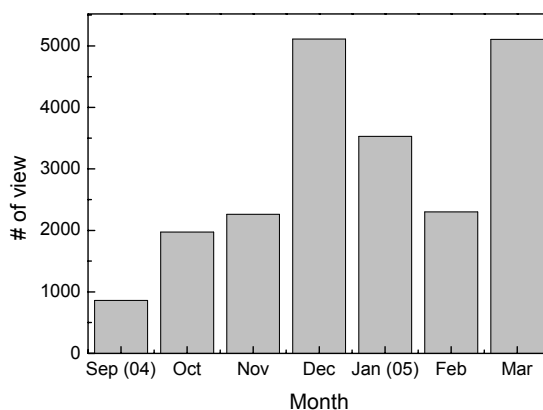


图 2、用户查看规则集统计（按 IP）

图 3 就是在 Unix/Linux 服务器上使用的用户统计（按不同 IP），其深灰色表示老客户，即上个月已经出现的 IP。

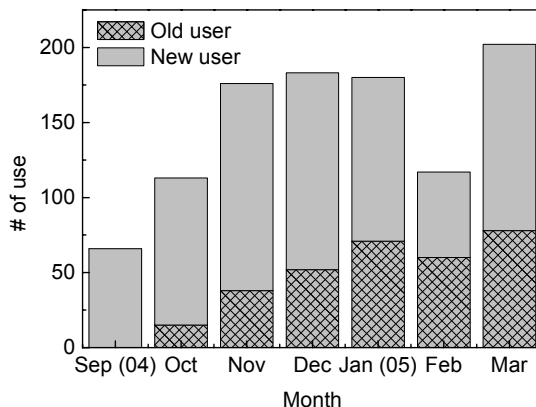


图 3、用户使用规则集统计（按不同 IP）

1.7 Chinese_rules.cf 的使用方法

下载 Chinese_rules.cf，把该规则放在 SpamAssassin 存放规则的目录（一般在 /usr/share/spamassassin）。通过 wget 下载的命令如下：

```
# wget -N -P /usr/share/spamassassin www.ccert.edu.cn/spam/sa/Chinese_rules.cf
```

每次更新 Chinese_rules.cf 都需要重启加载 SpamAssassin 规则的程序。如果你用 spamd 则通常重启的方法是：

```
# ps -ax | grep spamd
```

察看 spamd 进程的 PID，然后

```
# kill -HUP PID
```

如果你用 mimedefang 则要重起 mimedefang。假设 mimedefang 的重起脚本为 /etc/init.d/init-script，则命令如下：

```
# /etc/init.d/init-script restart
```

CCERT 每周更新一次规则集和相应分数，更新使用 CCERT 反垃圾邮件服务在 6 个月内处理过的垃圾邮件为样本。经常更新 Chinese_rules.cf 会使过滤效果更好。只要把上述下载命令以及重起 mimedefang 的命令放在 crontab 中，并定期运行就可以完成自动更新功能。假如你想一个月更新一次，那么在 root 的 crontab 中应该添加一行：

```
0 0 1 * * wget -N -P /usr/share/spamassassin www.ccert.edu.cn/spam/sa/Chinese_rules.cf; /etc/init.d/init-script restart
```

更多信息请参见第 2、3 和 4 节。

2. 在 sendmail 系统过滤中文垃圾邮件

(本节内容主要参考文献[1])

2.1 框架

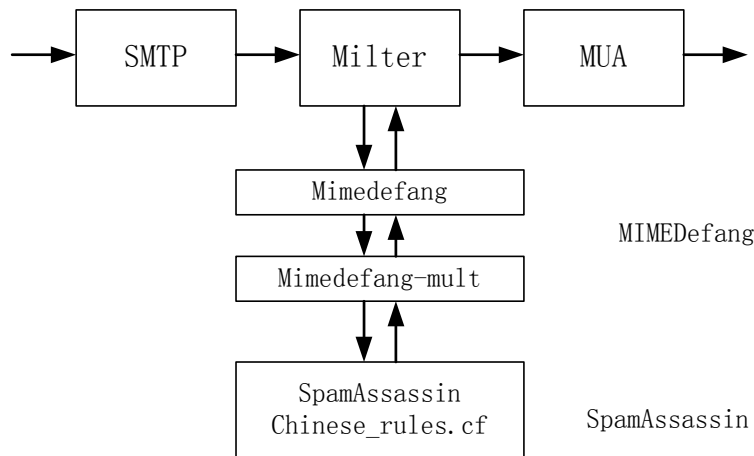


图 4、Sendmail 和 SpamAssassin 结合框架图

在 sendmail 的邮件服务器上安装 SpamAssassin 和 Chinese_rules.cf 的框架如图 4 所示，其中 SMTP、Milter 和 MUA 是 sendmail 的原始模块；其它都是需要安装的模块。在安装垃圾邮件处理功能之前，SMTP 每收到一封邮件就直接交给 MUA。在需要安装的组件中，MIMEDefang 就是 sendmail 和 SpamAssassin 之间的接口。每次收到一封邮件时，milter 模块就调用 MIMEDefang 模块，并将返回的检测结果送到 MUA。

2.2 安装 SpamAssassin

SpamAssassin 可以通过 CPAN 安装。这个方法很简单，只要进入超级用户权限然后键入如下命令：

```
# perl -MCPAN -e 'install Mail::SpamAssassin'
```

另一种安装方法是从 <http://spamassassin.apache.org> 网站上下载最新安装源代码，即“.tar.gz”文件，安装步骤如下（root 用户）：

```
# tar xzf Mail-SpamAssassin-2.64.tar.gz
# cd Mail-SpamAssassin-2.64
# perl Makefile.PL
```

```
# make
# make install
```

2.3 安装 Mimedefang

如果系统没有装 MIME::Tools 6 以上的版本, 你必须先安装这个 Perl 模块后才安装 mimedefang。安装 MIME::Tools 模块的命令如下:

```
# perl -MCPAN -e 'install MIME::Tools'
```

从 <http://www.mimedefang.org> 网站上下载最新的 mimedefang 安装源代码, 即 “.tar.gz” 文件, 安装步骤如下 (root 用户):

```
# tar xzf mimedefang-2.44.tar.gz
# cd mimedefang-2.44
# ./configure
# make
# make install
```

创建一个用户名字为 defang, mimedefang 运行的时候使用该用户的权限。

```
# adduser defang
```

在 mimedefang 源代码中有启动 mimedefang 的脚本: examples/init-script。把这个脚本拷贝到/etc/init.d 目录下。启动 mimedefang 的命令如下:

```
# /etc/init.d/init-script start
```

2.4 配置 Sendmail

在 sendmail.cf 文件中要添加两行, 第一行在 “# Input mail filters” 行下面添加 “O InputMailFilters=mimedefang” 如下:

```
# Input mail filters
O InputMailFilters=mimedefang
```

另外一个地方是在 MAIL FILTER DEFINITIONS 下面添加 “Xmimedefang, S=unix:/var/spool/MIMEDefang/mimedefang.sock, F=, T=S:60s;R:60s;E:60s”, 如下:

```
#####
#####
##### MAIL FILTER DEFINITIONS
```

```
#####
#####
#####
Xmimedefang, S=unix:/var/spool/MIMEDefang/mimedefang.sock, F=, T=S:60s;R:60s;E:60s
#
#####
#####
#####
##### MAILER DEFINITIONS
#####
#####
#####

重起 sendmail:

# /etc/init.d/sendmail restart
```

2.5 安装 Chinese_rules.cf

下载 Chinese_rules.cf，把该规则放在 SpamAssassin 存放规则的目录(一般在 /usr/share/spamassassin)。

如果你用 spamd 则需要重起 spamd。通常重起的方法是：

```
# kill -HUP `cat /home/spamd/spamd.pid`
```

如果你用 mimedefang 则要重起 mimedefang，重起的方法如下（参见安装 Mimedefang）：

```
# /etc/init.d/init-script restart
```

2.6 自动更新 Chinese_rules.cf

CCERT 每周更新一次规则集和相应分数，更新使用 CCERT 反垃圾邮件服务在 6 个月内处理过的垃圾邮件为样本。经常更新 Chinese_rules.cf 会使过滤效果更好。你可以用 wget 脚本去下载：

```
# wget -N -P /usr/share/spamassassin www.ccert.edu.cn/spam/sa/Chinese_rules.cf
```

只要把上述下载命令以及重起 mimedefang 的命令放在 crontab 中，并定期运行就可以完成自动更新功能。假如你想一个月更新一次，那么在 root 的 crontab 中应该添加一行：

```
0 0 1 * * wget -N -P /usr/share/spamassassin www.ccert.edu.cn/spam/sa/Chinese_rules.cf;
/etc/init.d/init-script restart
```

2.7 注意

中文邮件老触发 SpamAssassin 缺省安装的一些规则导致误判，你最好把这些规则停用，方法是在 `/etc/mail/spamassassin/sa-mimedefang.cf` 文件中添加：

```
score BODY_8BITS 0
score CHARSET_FARAWAY 0
score CHARSET_FARAWAY_HEADER 0
score HTML_CHARSET_FARAWAY 0
score MIME_CHARSET_FARAWAY 0
score UNWANTED_LANGUAGE_BODY 0
```

3. 在 qmail 系统过滤中文垃圾邮件

(本节内容主要参考文献[1])

3.1 框架

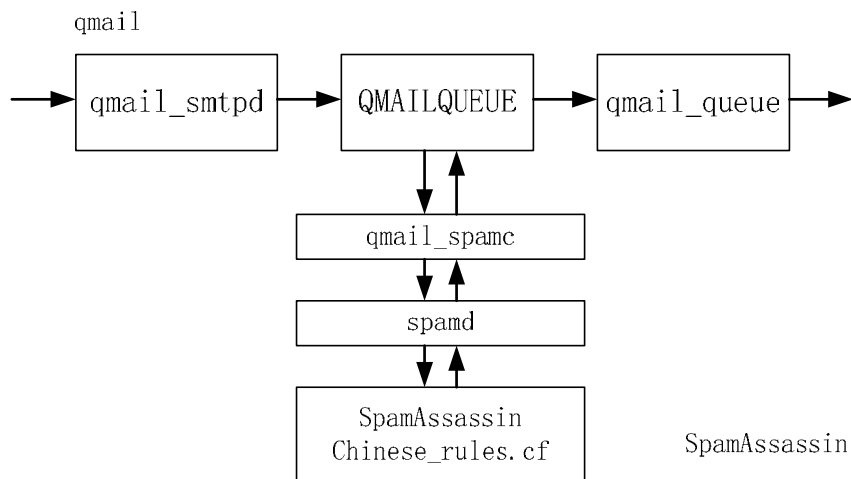


图 5、qmail 和 SpamAssassin 结合框架图

在 qmail 的邮件服务器上安装 SpamAssassin 和 Chinese_rules.cf 的框架如图 5 所示，其中 qmail_smtpd 和 qmail_queue 是 qmail 的原始模块；QMAILQUEUE 是 qmail 的一个补丁程序。其它都是需要安装的模块。在安装垃圾邮件处理功能之前，qmail_smtpd 每收到一封邮件就直接交给 qmail_queue。在需要安装的组件中，qmail_spamc 就是 qmail 和 SpamAssassin

之间的接口。每次收到一封邮件时，QMAILQUEUE 模块就调用 `qmail_spamc`，并将返回的检测结果送到 `qmail_queue`。

3.2 安装和配置 qmail

如果 qmail 已经安装和配置则可以跳过这一步。

按照 David Sill 编写的“Life with qmail” (<http://www.lifewithqmail.org>) 文档来安装和配置 qmail。其中需要安装 qmail, ucspi-tcp 和 daemontools。

3.3 安装和配置 SpamAssassin

SpamAssassin 可以通过 CPAN 安装。这个方法很简单，只要进入超级用户权限然后键入如下命令：

```
# perl -MCPAN -e 'install Mail::SpamAssassin'
```

另一种安装方法是从 <http://spamassassin.apache.org> 网站上下载最新安装源代码，即“tar.gz”文件，安装步骤如下（root 用户）：

```
# tar xzf Mail-SpamAssassin-2.64.tar.gz
# cd Mail-SpamAssassin-2.64
# perl Makefile.PL
# make
# make install
```

3.4 安装 Chinese_rules.cf

下载 Chinese_rules.cf (http://www.ccert.edu.cn/spam/sa/Chinese_rules.cf)，把该规则放在 SpamAssassin 存放规则的目录（一般在 `/usr/share/spamassassin` 或者 `/usr/local/share/spamassassin`）。

3.5 qmail 与 SpamAssassin 结合

在 SpamAssassin 3.x 的 `spamc` 目录下编译 `qmail-spamc`

```
# cc -O -o qmail-spamc qmail-spamc.c
# install -m 755 qmail-spamc /var/qmail/bin
```

确保 `qmail-queue`, `spamc` 和 `spamd` 在缺省路径中

```
# ln -s /var/qmail/bin/qmail-queue /usr/bin/qmail-queue
```

```
启动 spamd
/usr/bin/spamd --daemonize --pidfile /var/run/spamd.pid
每次更新 Chinese_rules.cf 需要重起 spamd 方法如下
# kill -HUP `cat /var/run/spamd.pid`
```

增加 qmail-smtpd 运行需要的内存空间
编辑/var/qmail/supervise/qmail-smtpd/run 中 softlimit 的-m 参数，一般 10M 就可以。

```
编辑 /etc/tcp.smtp 如下
127.:allow,RELAYCLIENT=""
:allow,QMAILQUEUE="/var/qmail/bin/qmail-spamc"
```

```
使所有的邮件都经过 SpamAssassin 处理
# qmailctl cdb
```

这样，所有的邮件都经过 SpamAssassin 处理，在邮件信头会加上一些扩展信头 X-Spam

4. 在 Windows 系统过滤中文垃圾邮件

(本节内容主要参考文献[1])

4.1 框架

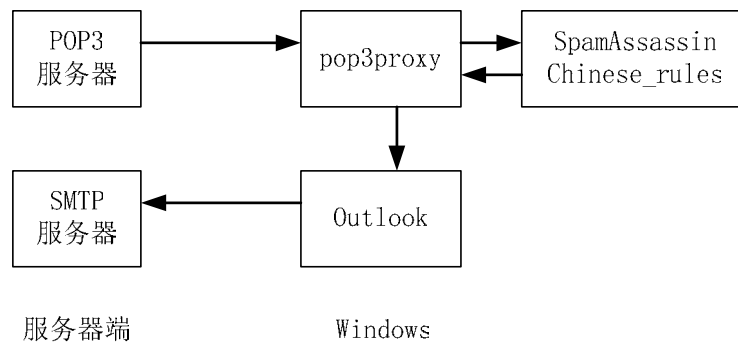


图 6、在 Windows 系统安装 SpamAssassin 的框架

在 Windows 系统上安装 SpamAssassin 和 Chinese_rules.cf 的框架如图 6 所示，假设 Windows 用户使用 Outlook，经过 SMTP 服务器发送信件和经过 POP3 服务器接受邮件。我们需要安装 pop3proxy、SpamAssassin 和 Chinese_rules.cf。Outlook 通过 pop3proxy 接收邮件，pop3proxy 每次收到一封邮件时就调用 SpamAssassin，并将返回的检测结果送到 Outlook。

4.2 安装 pop3proxy

从 <http://mcd.perlmonk.org> 网站上下载 pop3proxy.zip, 解压到 C:\pop3proxy 目录下。

从 <http://www.activestate.com/Products/ActivePerl> 网站下载 ActivePerl 5.8.3(注意有的其他版本不支持), 安装到 C:\perl 目录下。

运行 Perl Package Manager 程序 (刚安装), 并在 ppm> 键入 install Time::HiRes。用 quit 命令推出。

4.3 安装 SpamAssassin

从 <http://spamassassin.apache.org> 网站上下载 SpamAssassin 2.64 版本 (注意 3.00 版本不兼容)。解压缩。把 SpamAssassin 的 lib 目录中的所有东西拷贝到 C:\perl\site\lib。把 rules 目录拷贝到 C:\pop3proxy。

4.4 安装 Chinese_rules.cf

下载 Chinese_rules.cf 到 C:\pop3proxy\rules。注意下载后的名字应该是 Chinese_rules.cf。

4.5 配置

假设你有 2 个 POP3 服务器, 分别为 pop3.example1.net 和 pop3.example2.net。在 C:\pop3proxy 目录中创建一个文本文件名字为 hostmap.txt。该文件的内容是:

```
9110 = pop3.example1.net:110
8110 = pop3.example2.net:110
```

同时,你要在收邮件的软件上设置,以 Outlook Express 为例, 设置的方法如下:

工具-> 帐户

选择 pop3.example1.net 的属性, 在“服务器”的接收邮件 (POP3) 设成: 127.0.0.1, 在“高级”的接收邮件 (POP3) 设为 9110。

选择 pop3.example2.net 的属性，在“服务器”的接收邮件（POP3）设成：127.0.0.1，在“高级”的接收邮件（POP3）设为 8110。

4.6 自动启动 pop3proxy

启动 pop3proxy.pl 的命令如下：

```
c:\perl\bin\wperl c:\pop3proxy\pop3proxy.pl
```

如果你想开机时自动启动，那么只要在注册表（regedit）中的 HKCU-> Software-> Microsoft-> Windows-> CurrentVersion-> Run 中新建字符串值为上述的启动命令。

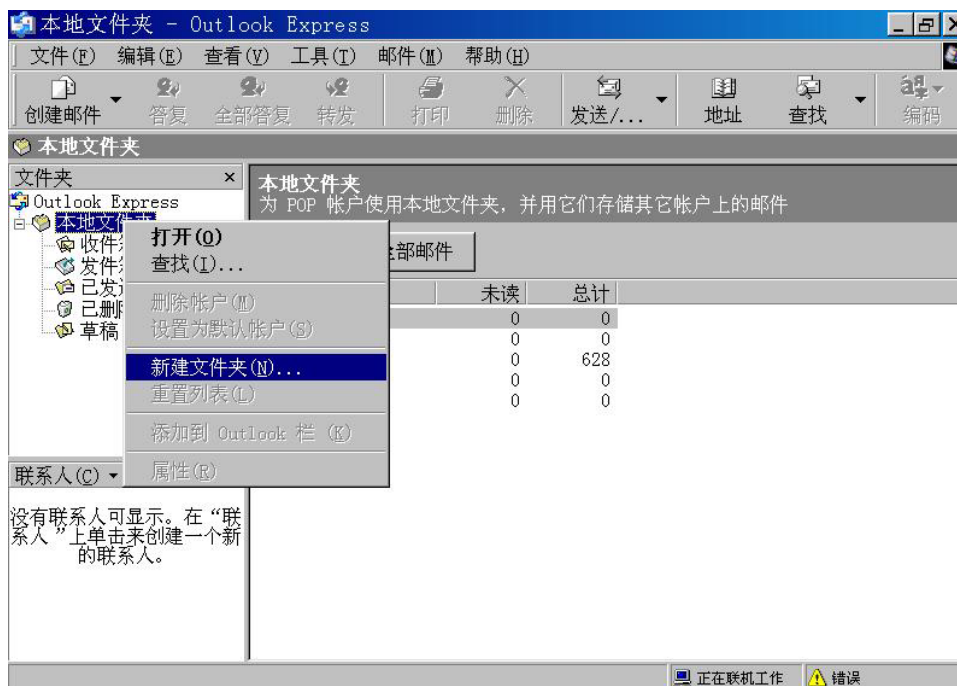
4.7 自动更新 Chinese_rules.cf

CCERT 每周更新一次规则集和相应分数，更新使用 CCERT 反垃圾邮件服务在 6 个月内处理过的垃圾邮件为样本。经常更新 Chinese_rules.cf 会使过滤效果更好。

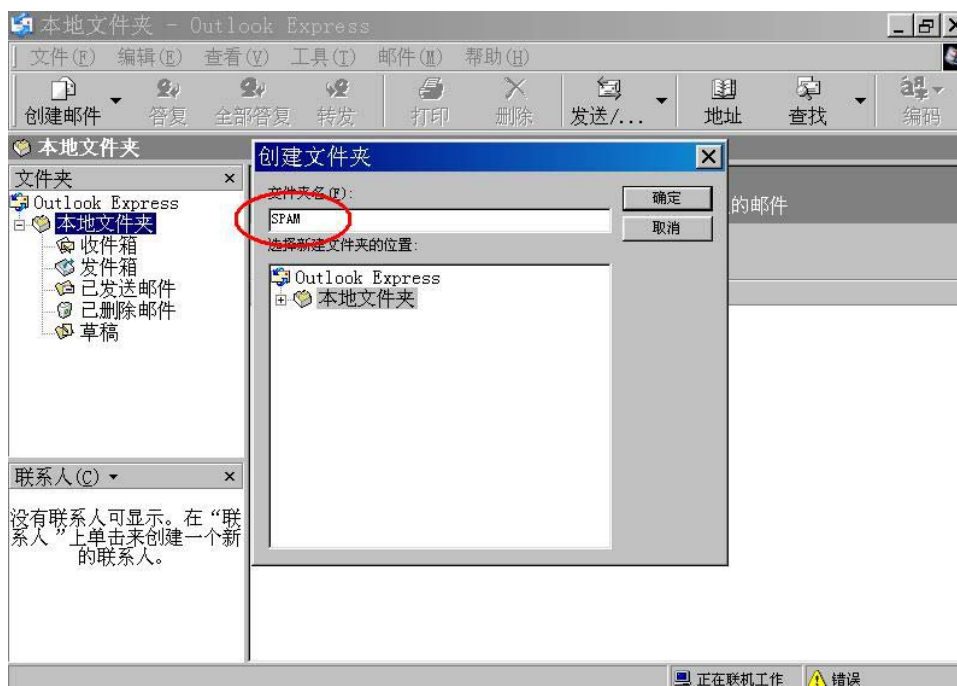
5. Outlook Express 设置步骤

5.1 建立一个垃圾邮件文件夹名字为 SPAM

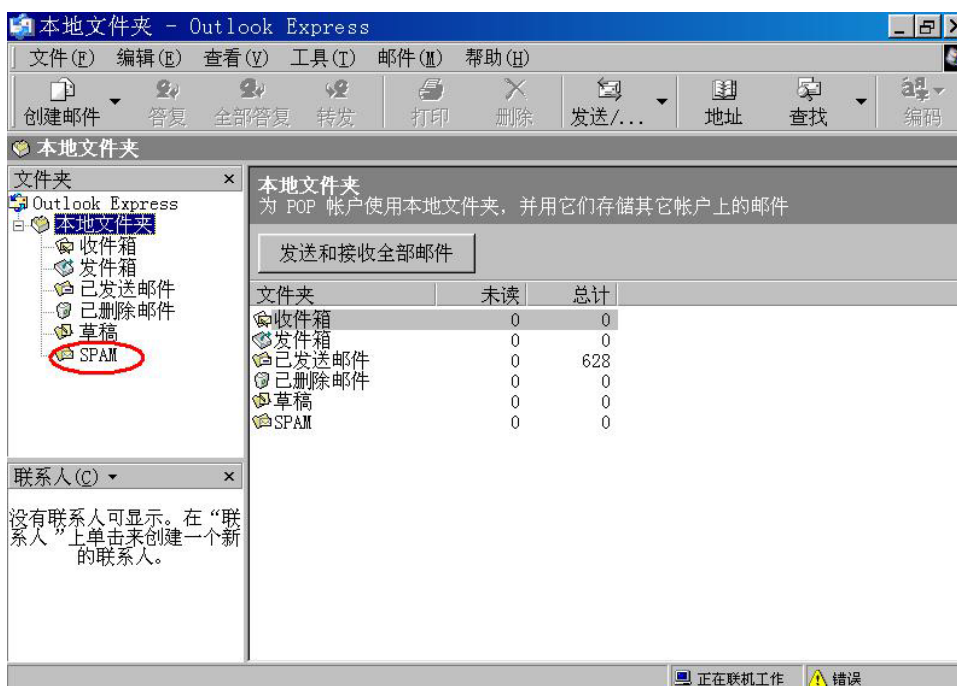
把鼠标放在“本地文件夹”上，按右键，再弹出窗口中选择“信件文件夹”：



在“文件夹名”的文本输入框内键入垃圾邮件文件夹名字为“SPAM”，单击“确定”。

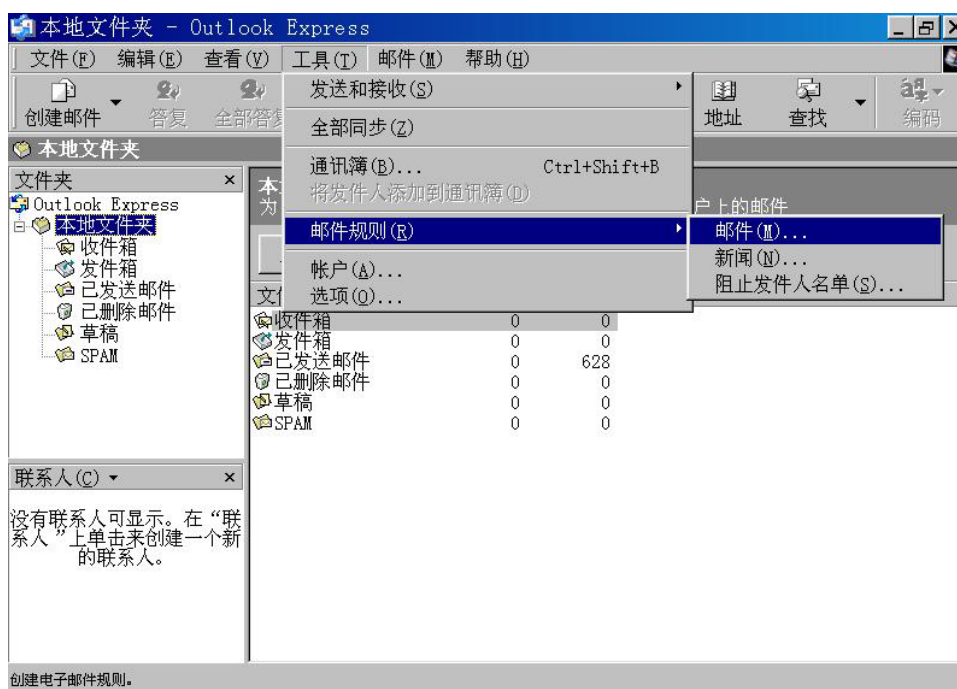


你会看到刚创建的文件夹 SPAM:

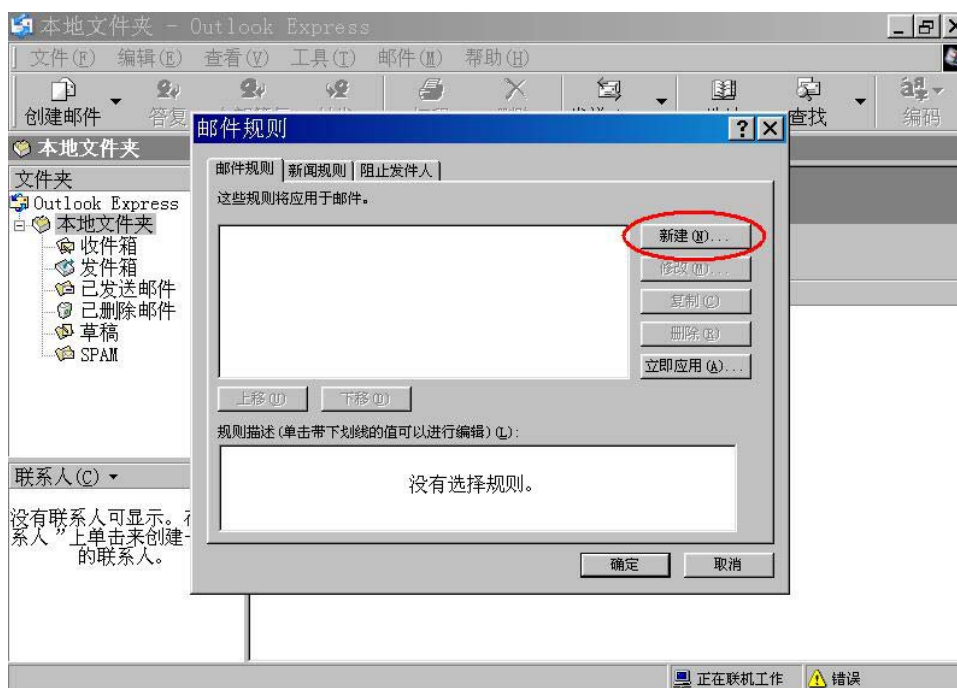


5.2 添加邮件规则把垃圾邮件移到文件夹 SPAM 中

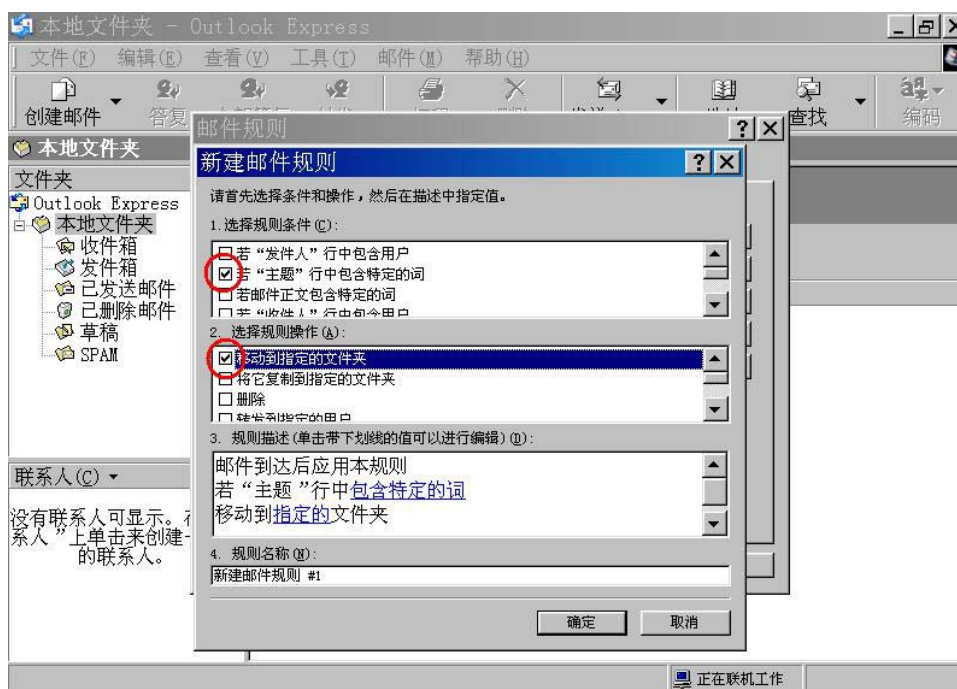
在主菜单选择工具 -> 邮件规则 -> 然后再选择邮件。



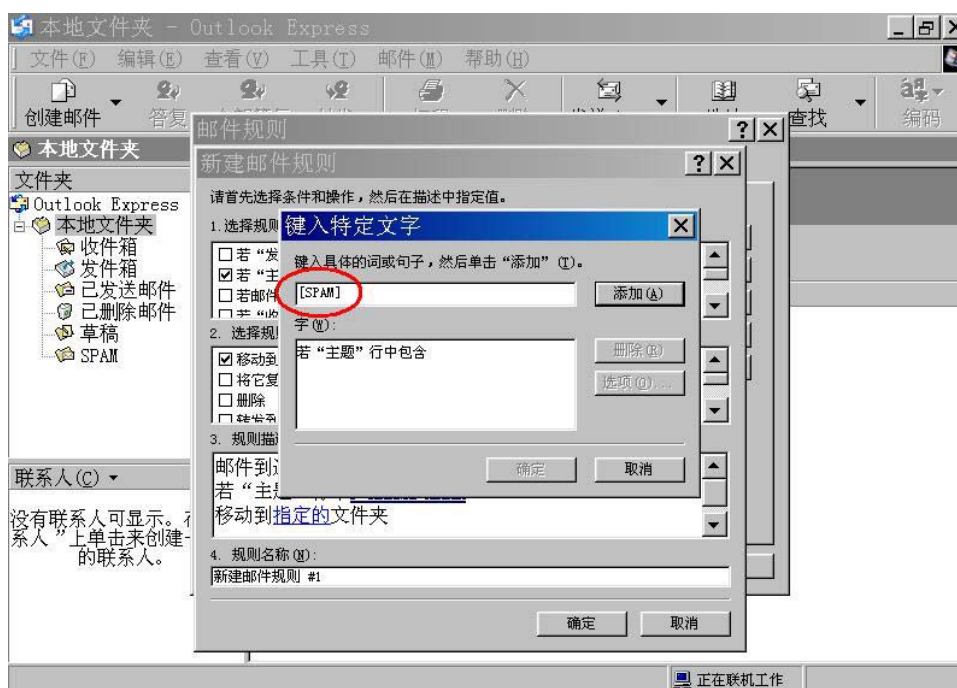
点击“新建”：



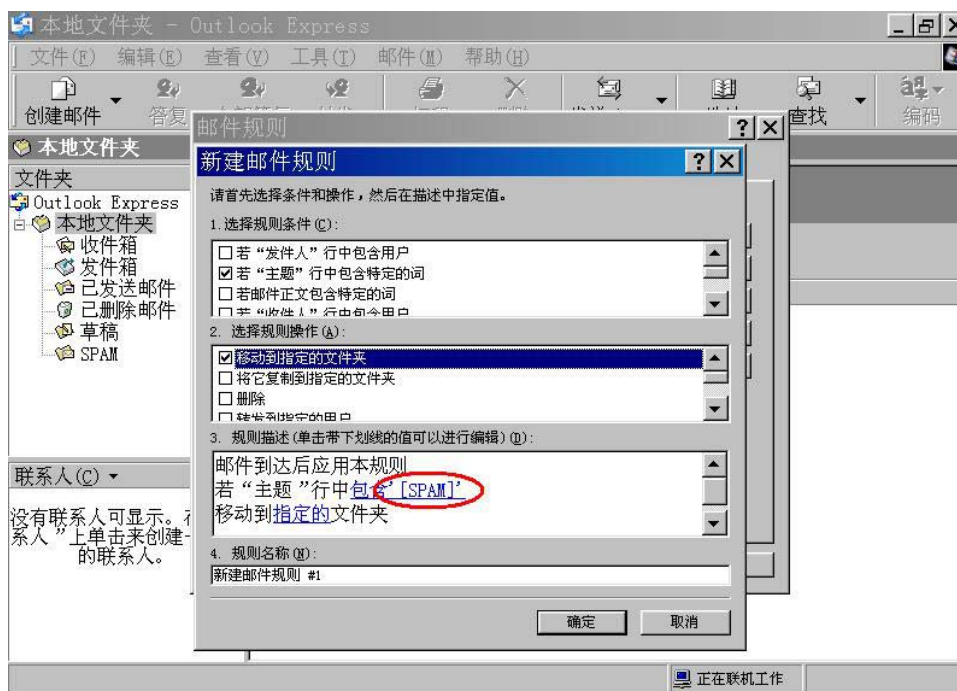
规则条件选择‘若“主题”行中包含特定的词’，规则操作选择‘移动到指定的文件夹’。在规则描述框中单击“包含特定的词”。



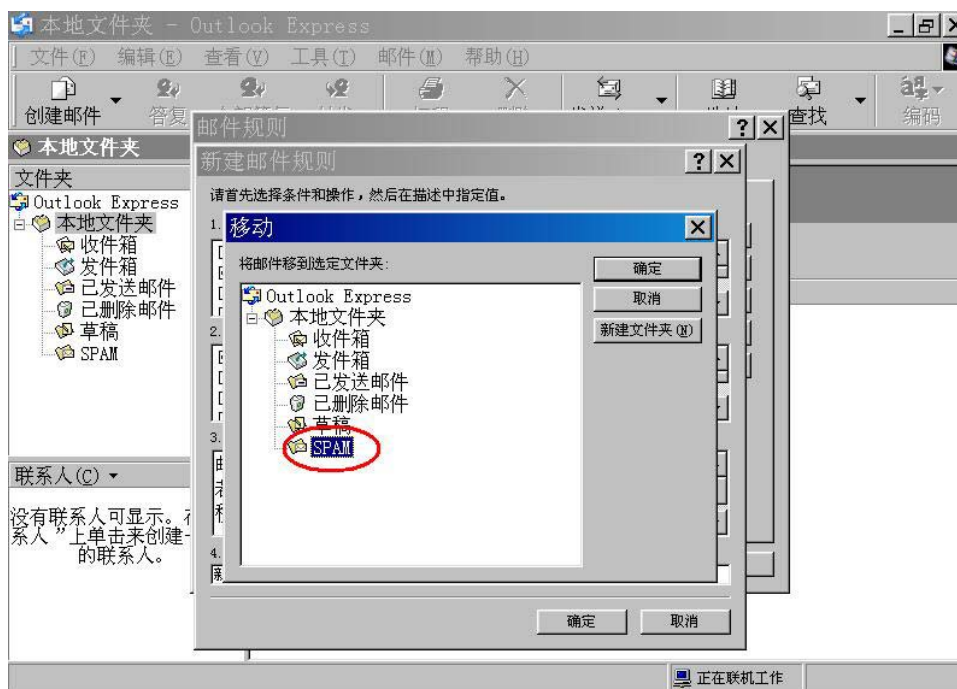
键入 “[SPAM]”，然后单击“添加”。



返回到原来的窗口出现“包含特定的词”被改为“包含[SPAM]”。在规则描述框中单击“指定的”

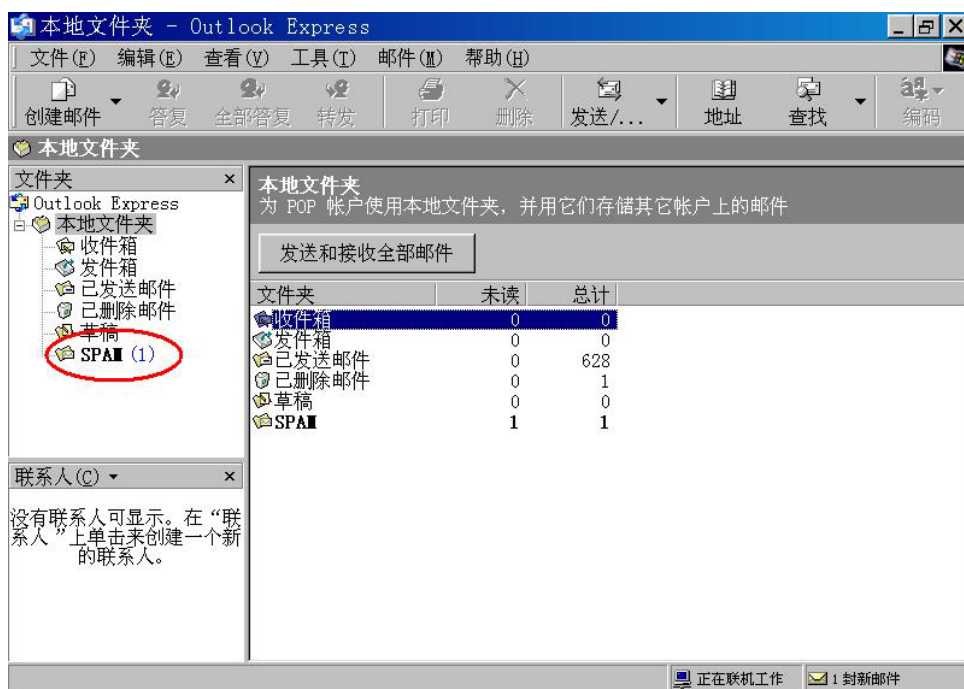


选择 SPAM 文件夹，单击“确定”。

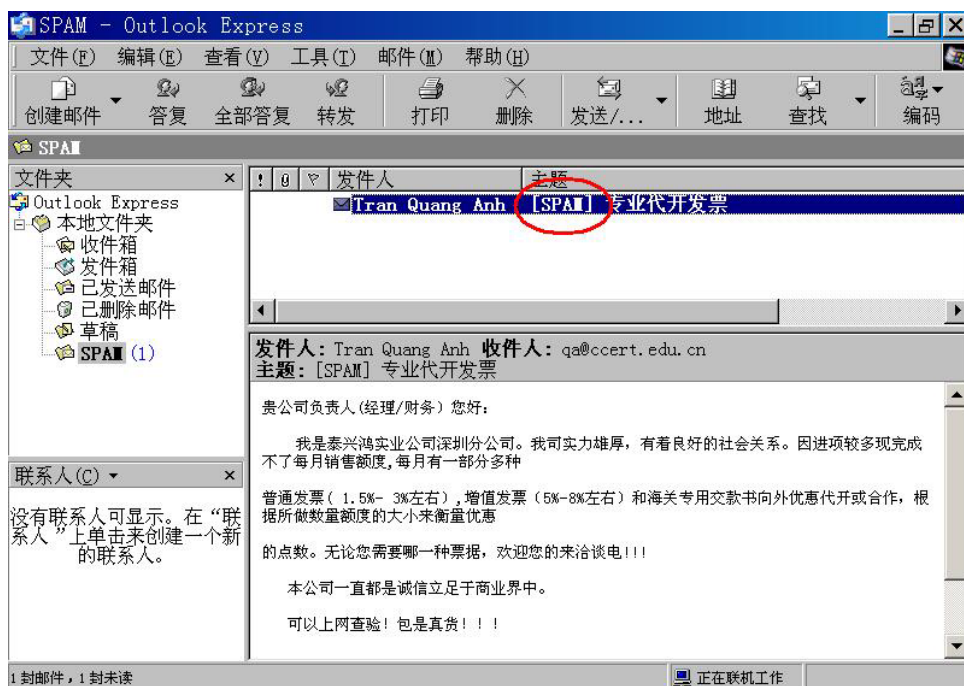


5.3 察看垃圾邮件

收到一封垃圾邮件时，Outlook 会自动把邮件放在 SPAM 目录下，并在文件夹 SPAM 上标志有新的邮件。



单击文件夹 SPAM 可以看到垃圾邮件的标题前面加上[SPAM]标志。



6. 参考文献

- [1] Schwartz, SpamAssassin, O'Reilly, 2004
- [2] SpamAssassin, <http://spamassassin.apache.org>
- [3] CCERT, <http://www.ccert.edu.cn>
- [4] Chinese_rules.cf, http://www.ccert.edu.cn/spam/sa/Chinese_rules.cf
- [5] Q.A. Tran, Statistical Chinese rules for SpamAssassin, <http://ccas.org.cn/lecture.html#3>
- [6] Mimedefang, <http://www.mimedefang.org>
- [7] David Sill, Life with qmail, <http://www.lifewithqmail.org>
- [8] Pop3proxy, <http://mcd.perlmonk.org/pop3proxy/>