一种贝叶斯方法邮件过滤的实现【初稿】

顾昊 钱晓俊
(中国科学院软件研究所，信息安全技术工程中心，北京，100080)
(中国科学院研究生院，北京，100080)

一、 背景

随着不断增多的用户接入互联网，电子邮件（E-mail）迅速成为最便捷和经济的交流方式之一。由于发送电子邮件非常容易、成本又非常之低，使得电子邮件成为一种电子化的手段被人利用，他们一般具有某种目的性地大量发送电子邮件。本文中，我们称自动生成的、不请自来的邮件为垃圾邮件（SPAM）[1]，垃圾邮件一般带有商业性。

近年来，垃圾邮件的数量呈指数趋势增长，据统计，2003 年收发的所有邮件中，约有 75%是垃圾邮件。大量的垃圾邮件不但浪费邮件者的时间，而且极大消耗的网络传输资源、邮件服务器的存储空间。由于垃圾邮件问题的严重性，目前，各大邮件服务提供商或者邮件客户端都提供了垃圾邮件过滤功能，起到了一定的作用。但是，这些自动过滤方法一般都需要用户自己输入有效的过滤规则；更重要的是，垃圾邮件本身具有易变性，这就要求用户不断改进和完善规则。规则的定义和修改是耗时和乏味的工作，并且对于用户来说，极易制定不恰当的规则。

二、 贝叶斯理论

要解决这一问题，必须建立一个既具有自适应性又能够个性化的自动邮件过滤系统。如何识别自动识别垃圾邮件呢？对接受者来说, 垃圾文件是一目了然的。若你雇佣某人来读你的邮件并把垃圾邮件删除, 他几乎不会有什么难题。可在缺少人工智能的情况下, 我们应做些什么来模拟这项工作呢？

贝叶斯方法非常适合建立这一系统，我们可以用相当少的算法来解决这个问题。事实上，我们只需要把个别单词的垃圾可能性找出来，进行一种贝叶斯组合，就可以很好的过滤垃圾邮件。而且贝叶斯方法是根据邮件内容动态调整的，新的垃圾邮件和正常邮件会不断调整邮件内容的垃圾邮件概率，从而适应每一个用户的需求。

贝叶斯定理： $P(H \mid E,c) = \dfrac{P(H \mid c) * P(E \mid H,c)}{P(E \mid c)}$

贝叶斯理论总体来讲是相当简单的：通过对某一事件过去发生概率情况的考查，大致可以推断出当前这一事件的发生概率。关于贝叶斯理论的详细介绍，读者可以查阅参考文献[5]。我们只给出这样一个结论，通过对考查大量垃圾邮件样本中各种因素的概率情况，在得到一封未知邮件时，通过分析这封邮件中各

个因素的情况，大致可以推断出这封邮件是否为垃圾邮件。

根据贝叶斯理论，我们开发了垃圾邮件过滤系统 AntiSpam，经测试，能较好的解决上面的问题。下面我们将分别介绍 AntiSpam 系统的情况以及贝叶斯方法的实现原理。

三、 AntiSpam 系统介绍

AntiSpam 是我们使用 Java 开发的服务器端邮件接受系统，目前支持 POP3 协议（IMAP 协议尚未开发）从其他邮件服务运营商处取电子邮件，收取邮件的同时，完成对邮件的自动分类，判断该邮件是否为垃圾邮件。该系统可以根据用户的反馈进行更新，从而更好的运用贝叶斯方法计算垃圾邮件的概率。

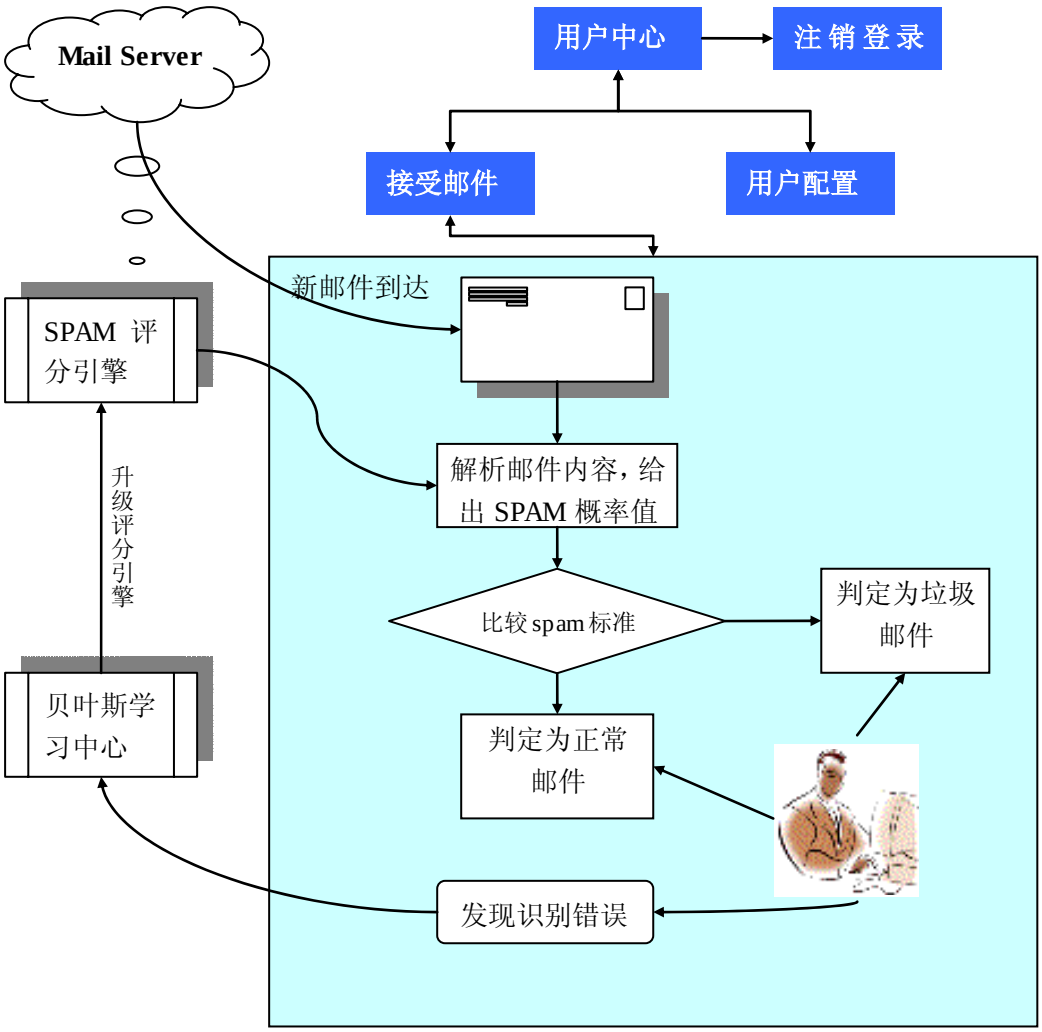

图 1 系统工作流程图

图 1 给出了系统的工作流程图，描述了系统的总体工作情况。下面我们分步骤介绍用户如何使用本系统的情况。

用户登录后，可以进入配置中心（如图 2 所示），用户在这里可以添加、删除、修改 Pop3 服务信箱参数。配置完毕后，用户进入收件箱，查看邮件列表（如

图 3 所示)，特别指出的是，其中状态一栏，是通过贝叶斯评分引擎计算后，再根据系统标准，识别出的 SPAM/HAM 状态。用户点击邮件标题链接，可以查看邮件的正文，如图 4 所示。

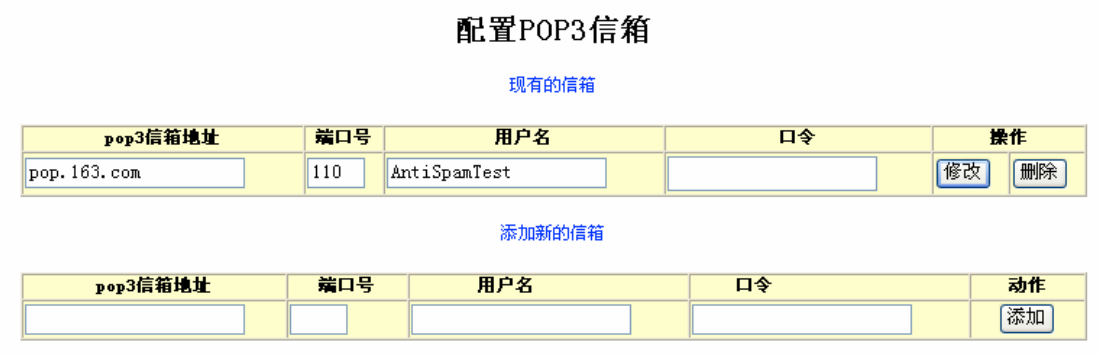

图 2 用户配置中心图示

图 3 查看邮件列表图示

图 4 查看邮件正文

如果用户发现系统识别错误，可点击图 4 中的[提交]按钮。系统将接受用户反馈，通过贝叶斯学习中心，改进评分引擎，使得系统更好得为用户服务。

出于系统运行效率的考虑，AntiSpam 并不及时相应用户的每一个反馈请求，而是首先收集用户的反馈，分别保存误判为垃圾邮件、没有被识别出的垃圾邮件的样本，最后统一升级评分系统。这样做的目的还处于另一种考虑，系统可以对用户的反馈作出进一步分析，避免用户的恶意反馈。

四、 实现原理

整个 AntiSpam 的实现机制得益于 Bayesian 定理之不确定性推理模型的帮助和 Paul Graham 之详细说明的启发（见参考文献[1]，[2]）。Bayesian 定理是一种利用概率论识别 spam 的方法，它的特点在于并不想去抽取 spam 的个体性质，而是统计邮件中各个 token 可能导致它为 spam 的概率，然后利用 Bayesian 定理计算该邮件为 spam 的可能性。

Spam 过滤的第一步是分析一定数量的垃圾邮件和正常邮件语料，建立起两个散列表。分析一份邮件为单独的 token 集合，这里把空格、逗号、句号、分号和阿拉伯数字当作分隔符，其他的都当作 token。把 token 和 token 出现的次数建立 spam hash 和 ham hash。

当完成了两张 hash 表，第二步就是根据它们建立最重要的 hash 表，称为 prob hash，这张表记录一个 token 可能引起邮件为 spam 的概率。首先，获取任意的邮件，分解该邮件字符串为 token，根据 spam hash 和 ham hash 计算这些 token 可能导致一份邮件被判断为 spam 的概率。计算方法如下：

假设准备的 token 串为：$t1,t2,...,tn$，它们在 spam hash 中出现的次数为：$N_{t1},N_{t2},...,N_{tn}$，在 ham hash 中出现的次数为：$M_{t1},M_{t2},...,M_{t3}$，spam hash 中 token 的次数之合为 $NUM(spam)$，ham hash 中 token 的次数之合为 $NUM(ham)$。于是容易得到 $t1,t2,...,tn$ 在 spam 和 ham 中出现的概率：

$P1(ti)=\frac{N_{ti}}{NUM(spam)}$，token ti 在 spam 中出现的概率。

$P2(ti)=\frac{M_{ti}}{NUM(ham)}$，token ti 在 ham 中出现的概率。

若设事件 A 为 token ti 出现的那份邮件为 spam，计算那些可以写入 prob hash 的概率，建立 prob hash：

$$P(A|ti)=\frac{P1(ti)}{P1(ti)+2P2(ti)}$$

这里的一些计算细节如下：

- 为了防止 prob hash 这张散列表扩展得太快，不计算那些在 ham hash 和 spam hash 中出现次数不超过 3 次得 token。
- 为了降低过滤系统的纠错率，把在 ham 中出现的 token 的概率加倍，这样可以避免那些偶尔在 ham 中出现，或者从不出现的 token，使得邮件被错判。
- 为了控制一个 token 不至于过分影响整个概率的计算，限制计算得到的概率 $0.01\le P(A|ti)\le 0.99$。

接下来，就可以根据这个 prob hash 来判断一份邮件是否为 spam。由 Bayesian

定理，计算 token 可能导致邮件为 spam 的概率。当一份待检查的邮件 M 到达的时候，同样根据确定的 delimiter 把邮件 M 的邮件体分解为 token 的集合：$\{t1,t2,...,tn\}$，这里只挑选最关键的 15 个 token，当然，若是只想挑选 10 个也没有问题。这样做的好处可以降低处理需要的时间，而且更容易发现问题的结症。关键程度的确定就看该 token 概率值与 0.5 这个中间值的距离。当一个 token 没有在 prob hash 中出现，设定该 token 的概率值为 0.4，因为若是一个 token 几乎既没有在 spam hash 中出现过，也没有在 ham hash 中出现过，那么大抵这个 token 是个正常的 token。这样我们可以得到 $\{t1,t2,...,tn\}$ 的最关键的 15 个概率值：$p1,...,p15$，则最终该邮件 M 为 spam 的概率：

$$P(A|t1,...,tn)=\frac{\prod_{i=1}^{15} pi}{\prod_{i=1}^{15} pi+\prod_{i=1}^{15}(1-pi)}$$

于是得到最终结果。

最后要说明的就是工作中的学习，当发现一份邮件被纠错。可能是 spam 被漏过，或者 ham 被误判。以 ham 被误判为例，那么我们就把该邮件 M 中的 token 在 ham hash 中出现的次数增加一个额度，该额度可以为该次数的倍数之类。因为当被误判，说明该 token 在 ham hash 中被重视的程度不够，若是 spam 被漏过，则操作类似。然后，我们再根据 spam hash 和 ham hash 重新计算出新的 prob hash，这样学习过程结束。

五、 结论

这里给出一组测试数据，规模在 2000 左右，分别列出 SPAM 识别率和 HAM 识别率。并简要解释误判的原因。测试语料跟建立数据语料不重合，测试数据如下：

当中间值设置为 0.5 的时候：

	学习	测试	识别结果	识别率
Spam	432	49	43	87.7%
Ham	1447	242	242	100%

当中间值设置为 0.45 的时候：

	学习	测试	识别结果	识别率
Spam	432	49	46	93.3%
Ham	1447	242	240	99.2%

下边分析一下漏掉的三份邮件的特点：

Spmsgc62.txt

Subject: view the hollander collection

view the hollander collection t h e h o l l a n d e r c o l l e c t i o n f i v e a r t i s t s . o n e f a m i l y gino hollander . painting jim hollander . photography siri hollander . sculpture to view : http : / / www . hollanderart . com scott hollander . photography barbara hollander . writing

Spmsgc83.txt

Subject: junk mail : books for linguists

plurabelle books has a new catalogue of second hand and out of print books in linguistics and history of linguistics available . it contains 250 titles . please ask for your free copy of the catalogue (please include your mailing address if you want a paper copy) or visit us on the net http : / / www . plurabelle . co . uk / lingu . htm or write to dr michael cahn plurabelle books 77 garden walk cambridge cb4 3ew tel 0044 - 1223 - 366680 , fax - 571105 lingu @ plurabelle . co . uk we only sell books we would like to read open anytime at http : / / www . plurabelle . co . uk

Spmsgc90.txt

Subject: webmining

free white paper on data mining web data : http : / / www . webminer . com / paper . htm

这里是统计出来的它们关键 token 以及关键 token 所代表 spam 的概率：

Spmsgc62.txt

y : 0.05120838561530749
l : 0.07193476937829474
l : 0.07193476937829474
view : 0.10025733641911937
writing : 0.1295031048785931
e : 0.1878470849407081
m : 0.20397250900657546
com : 0.7082711898284206
www : 0.3212555497412582
h : 0.0773922374555034
e : 0.1878470849407081
h : 0.0773922374555034
o : 0.19801969679251974
l : 0.07193476937829474
l : 0.07193476937829474

Spmsgc83.txt

anytime : 0.99
junk : 0.9456016443343837
dr : 0.04241979895164473
walk : 0.07746984253485818
@ : 0.08626923582658803
uk : 0.09592833749677
uk : 0.09592833749677
books : 0.13524091834765078
books : 0.13524091834765078
co : 0.22050802250247348
would : 0.26130198054406384
of : 0.17769891909541274
second : 0.08667831391080784
read : 0.6400025521062006
at : 0.28308783194700937

Spmsgc90.txt

paper : 0.11104681852008605
htm : 0.5270277648690014
white : 0.4

```
paper : 0.11104681852008605
on : 0.3086366668823493
data : 0.04251316998343358
mining : 0.4
web : 0.6129813953501898
data : 0.04251316998343358
http : 0.47664738674024026
/ : 0.44360160656082376
/ : 0.44360160656082376
www : 0.3212555497412582
webminer : 0.4
com : 0.7082711898284206
```

分析一下它们共同的特点在于：

- 都很短小，未能出现 spam hash 中举足轻重的一些词汇
- 三份中有两份都包含着大量空格

而对于这两个特点，过滤器作出的反应是：

- 因为都很短小，一些关键词汇没有命中，因而得以逃脱。不过这点可以通过增加 spam hash 的长度得以改善，因为现在的过滤器 spam hash 只是通过 400 多份 spam 建立起来的，当这个数字增加到 1000，甚至 3000 的时候，我想过滤器的效果将会得到很大的改善。这点可以从对 ham 的纠错率是如此之低，可以看的出来，因为 ham hash 是通过将近 1500 份 ham 建立起来的，这个时候已经能够比较完美的体现 bayesian 的威力。
- 这里的缺陷在于 token 的 parse 方法。因为现在的过滤器使用简单的空格、逗号、句号、分号、冒号和阿拉伯数字来分词，这样将会导致产生很多的单个字母或符号的 token 出现，尤其当 message 很短小的时候，这种干扰更加明显。而且这样的问题在于没有进行 url 的识别。一个方法，就是改进 token 的分析方法，使得 token 的提取更加具有适应性。

六、 进一步的工作

从上面的介绍可以看出，AntiSpam 运用贝叶斯原理，可以较为满意的识别出垃圾邮件，误判率也很低。同时，AntiSpam 还存在很多可以改进的地方，下面列举一二：

- 没有进行 url 识别
- 没有区分主题和邮件内容的不同重要性及其相关联系
- 多语言支持（中文垃圾邮件识别）
- HTML 代码过滤区分（舍弃不必要的 HTML 标识符，提取合适的标识信息）
- 更多维特征向量的选择（词组识别、标题与正文赋予不同权值等）

参考资料

[1] A Plan for Spam. http://www.paulgraham.com/spam.html
[2] Better Bayesian Filtering. http://www.paulgraham.com/better.html
[3] Ling-Spam Corpus. http://www.agueb.gr/users/ion/publications.html
[4] 王文杰，叶世伟. 人工智能原理与应用. 人民邮电出版社
[5] An Introduction to Bayesian Networks and their Contemporary Applications http://www.niedermayer.ca/papers/bayesian/index.html
[6] A Bayesian Approach to Filtering Junk E-Mail (1998). http://citeseer.ist.psu.edu/sahami98bayesian.html