专题研究报告（一）

垃圾邮件的综合过滤方法

姓　　名：徐　激

学　　号：030963

导　　师：龚　俭

垃圾邮件的综合过滤方法[1]

徐激*　　龚俭

（东南大学计算机系，江苏 南京，210096）

（江苏省计算机网络技术重点实验室）

【摘要】本文研究了几种常用的垃圾邮件过滤算法，分析了它们在中文邮件环境中存在的问题。本文根据各算法的优缺点，将它们进行改进、叠加和相互结合，并通过查看发出的邮件内容进行辅助学习，从而建立一个垃圾邮件的综合过滤方法。文章最后对该综合方法的效率做了分析和比较。

【关键词】垃圾邮件；正常邮件；黑白名单；规则；贝叶斯过滤算法；

An Integrated Way to Filter Spam

Xu Ji　　Gong Jian

(Southeast University, Computer Science Dept., Nanjing, Jiangsu, P.R.China, 210096)

(JiangSu Network Technology Key Laboratory)

【Abstract】In this paper, several popular algorithms for filtering spam are studied, and some problems involved when they are employed in Chinese email environment are analyzed. This paper propose to improve, overlap, and join them together according to their advantage and disadvantage, and with the assistant study by checking the mails which are sent off, to set up an integrated way for filtering spam. The efficiency analysis and comparison is given at last.

【Key Words】Spam mail; Ham mail; White and Black Lists; Rule; Bayesian Filtering;

1. 引言

随着 Internet 的发展，电子邮件已成为一种常用的通讯方式。但由于其成本低廉、传播迅速，Internet 上出现了越来越多的不被请求的邮件，即垃圾邮件。这些不受欢迎的垃圾邮件使用户不得不花费大量的时间和精力来处理它们，从而严重影响了用户对电子邮件的正常使用。同时，垃圾邮件的过滤技术也在不断地发展，尤其是近几年出现了许多优秀的技术成果。但在中文环境中，这些过滤方法的效率均不够好，不能满足用户的要求。本文对几种常用的过滤算法进行了研究，分析了它们在中文环境中存在的问题，然后根据各算法的优缺点，对它们进行改进和相互结合，以叠加的方式对邮件进行多层过滤，并通过各算法之间自动传递辅助信息来提高算法本身的精确度。并且，本文提出了通过查看用户发出的邮件内容来进行辅助学习，从而提高自动获取知识的能力，并最终建立一个能够适应中文实际运行环境的综合的垃圾邮件过滤方法。最后，文章对该综合方法的效率进行了分析和比较。

2. 常用过滤算法

目前，有多种算法可以用来过滤垃圾邮件，并且已经在实际应用中发挥一定的作用，但它们本身又各有优缺点。

2.1 黑白名单算法（White and Black Lists）

这是最常用的一种算法。过滤系统在处理新到达的邮件时，首先查看邮件头部的发送方地址，对于地址处于白名单中的邮件将全盘接收，而对于处于黑名单中的邮件则直接拒收。该算法的优点是简单明确。但它有两个缺点：1）黑白名单在设定时必须准确。如果把友好地址列在了黑名单中，会造成误判。2）需要不断地更新和维护，并且通常无法涵盖所有的

[1] 本文受国家自然科学基金课题资助（90104031）

作者简介：徐激，女，硕士研究生，主要研究方向为网络安全；龚俭，工学博士，东南大学计算机系教授、博导，主要研究方向包括网络管理、网络安全、网络体系结构等。

情况。因此，黑白名单算法的效率并不高，由文献[2]中提供的数据表明，它只能过滤掉不超过 50%的垃圾邮件。

有一些著名的志愿者组织在他们的网站上维护着一系列的 IP 级的黑名单，它们或者是垃圾邮件发送者的地址，或者是那些具有严重安全漏洞的邮件服务器地址。任何 ISP 都可以订阅这些服务，使这类邮件在到达之前就自动被拒绝。目前比较值得信任的组织有 Spamhaus、dsbl、mail-abuse 以及中国反垃圾邮件联盟提供的 CBL 服务。

2.2 基于规则的过滤算法（Rule-Based Filtering）

该算法也称启发式算法，在 2002 年以前占据了主导地位。它是通过与既定的规则相比较来判定是否为垃圾邮件，这些规则包括：特别的词语，如“免费”、“订阅”等；伪造的信件头，如不合理的日期等；特别的格式，如红色粗体字等。可见，在该算法中，规则的制定和维护是至关重要的。如果选用像 SpamAssassin 这样优秀的工具，经过合适的参数调整，可以过滤 90%的垃圾邮件[2]。但是制定的规则总是静态的，很容易被垃圾邮件发送者们轻易的绕过。现有的工具已避免了单一的判断方式，而是通过一个阀值来衡量，当邮件与规则相匹配的总分值超过这个阀值，就会被定为垃圾邮件。

2.3 贝叶斯算法（Bayesian Filtering）

该算法也称统计算法，将它用于邮件过滤的想法最早是于 1998 年在美国人工智能联合委员会的一次会议上被 Pantel and Lin[4]和来自微软的一组成员[5]提出的，但是从文中提供的数据来看，它在实验中的效果不够理想，因此没有得到推广。2002 年 8 月，Paul Graham 在”A Plan for Spam”[1]一文中重新提出了这个想法，并用 Arc 语言实现该算法，取得了很好的效果，从而引起了广泛的关注。

贝叶斯算法的基本思想是通过对邮件头部和邮件信体中的单词进行概率计算，从整体上判断是否为垃圾邮件。单词的概率计算依赖于已知的垃圾邮件和正常邮件中单词出现的频率来完成，因此系统必须经过一段时间的学习和知识积累之后才能开始为用户工作。它的工作流程包括两个部分：1）学习过程。首先确定已知的垃圾邮件和正常邮件的集合，然后根据每个单词分别在两个集合中出现的次数，计算单词的垃圾概率。例如，在本文的实验中，单词“购买”的垃圾概率为 0.908，而“系统”的垃圾概率仅为 0.09；2）计算和判断过程。当一封新邮件到达时，系统需要对信件全部内容进行分词和选词，得到一组单词流，然后根据学习到的单词库中的信息，计算整个单词流的概率，并最终判断该信件是否为垃圾邮件。

自 2003 年以来，Internet 上相继出现了一些基于贝叶斯算法的邮件过滤工具，并且开始得到商业界的关注。Apple、Microsoft、AOL 等公司已经开始在他们的产品中采用该算法。贝叶斯算法得以如此迅速的兴起源于它优越的性能和良好的适应性。根据 Paul Graham 的叙述[3],他的工具可以过滤 99.5%的垃圾邮件，并且误报率低于 0.03%。贝叶斯算法是基于概率计算的，它不仅计算是否为垃圾邮件的概率，也计算是否为正常邮件的概率。例如，单词“购买”可以增强邮件的垃圾概率，而单词“系统”则会使它减弱。所以该算法对邮件做出的判断是总体上的，而不像基于规则的算法，仅仅由于邮件中出现几个特殊词语就简单地将其归为垃圾邮件。并且，贝叶斯算法是基于内容的，带有学习的性质，即知识会随着垃圾邮件内容的变化而更新。例如，垃圾邮件发送者常常会将英文单词“OK”写成“0K”来绕过规则检查的过滤器，而贝叶斯算法在学习时会发现它，并且“0K”的垃圾概率将会比“OK”的更高。所以，无论垃圾邮件发送者采用何种手段来逃避检查，邮件中总会包含不受欢迎的内容，而只要出现不受欢迎的内容就一定会被系统识别出来。

但是，要保持这种良好的性能，系统需要根据用户不断地反馈来获取知识，这也正是贝叶斯算法的一个缺点。当用户在系统无法做出判断，或判断失误时不能给予反馈，贝叶斯算法的效率会迅速下降；反之，如果用户不断地给系统提供反馈信息，系统则可以通过学习保持较高的性能。另外，贝叶斯算法是个性化的，易于单个用户的使用，因此综合算法中仍必

须保持这种个性化的特征。

2.4 其它算法

其它还有一些算法可以用于邮件的自动过滤。例如，签名算法（也称蜜罐算法）通过与已知的垃圾邮件做比较来判断，但这种算法只能过滤 50%-70%的垃圾邮件[2]；发送方响应算法，通过强制发送方做一些特定行为，如回答问题等，来阻止大量垃圾邮件的发送，但这种方式会严重影响正常邮件的使用。

3. 存在的问题

作者对中国教育科研网华东（北）地区网络中心的邮件服务器进行了观察分析。这个邮件服务器虽然只有几十个用户，但受工作性质的影响，平均每天的邮件量在上千封，垃圾邮件的数量很大，具有校园网邮件服务器的特征。通过观察可以发现，这个邮件服务器具有六大特征：

1) 邮件以中文为主，占所有邮件的 99%以上；

2) 垃圾邮件的数量很大，特别是一些对外公开的地址，单个地址一天就会收到数百封垃圾邮件。从总量上计算，垃圾邮件占到所有邮件的 93.1%；

3) 邮件服务的用户可分为三类：教师、工作人员、学生，其邮件内容也相对集中在这三类中；

4) 正常邮件中，有大量邮件来自于中国教育网内部，占所有正常邮件的 40.9%；而对于垃圾邮件，除了为数极少的伪造成教育网地址的垃圾邮件外，教育网内部几乎不产生垃圾邮件。数据表明，垃圾邮件中“From”地址为教育网地址的邮件占有率小于 0.1%，并且集中在几个相似的伪造地址上；

5) 中文邮件常以 Base64 编码传输；

6) 大量的垃圾邮件的“From”地址是一些免费邮箱的地址，如 163、263、yahoo 等。

作者将以上三种主要的过滤算法分别在这个环境中进行了测试，它们的效率都明显不高，甚至出现较高的误报率，因此不能被直接应用于实际的中文邮件环境。

在实验中选用的 dsbl 提供的黑名单只能过滤 5.4%的垃圾邮件。类似于 dsbl 的这些志愿组织所提供的黑名单中列举的是邮件服务器的 IP 地址，这种较底层的过滤方式过于粗糙。对于大型的邮件服务器来说，一旦 IP 地址被列入黑名单，该服务器的所有用户都会受到影响，因此大多数邮件服务提供商都会尽量避免被列入到这类名单中。本文的实验环境中，有大量的中文垃圾邮件来自于免费邮箱地址，它们在 IP 地址这一层是不能被识别出来的，此时黑名单算法的效率很低。

对于规则算法来说，其效率的高低取决于规则的制定。本文选用的 SpamAssassin 是一个常用的规则算法的工具，它的规则库中已经积累了六千多条规则，每一条规则都有默认的分值。但是这些规则是根据英文垃圾邮件的特征总结出来的，因此有很多规则并不适合中文邮件环境。并且，其阀值的制定是静态的，当阀值设定得较高时，系统的误报率降低，但漏报率明显增大，反之则会增加误报率，其性能不能满足用户要求。

贝叶斯算法虽然是最受瞩目的算法，但它仅在英文邮件环境中得到实现。对于中文邮件，目前还没有基于该算法的过滤工具被很好地设计实现出来。贝叶斯算法是一种基于信件内容的算法，它需要对出现的单词进行概率计算，从整体上判断邮件的性质。但是中英文在分词上存在巨大的区别：英文是用空格简单地分词，而中文的词与词之间没有直接的分词符号，通常是通过人的理解来划分的。因此，直接将该算法应用于中文环境显然是不合理的。本文选用 Spambayes 工具和双字分词法，通过修改该工具的源代码来增加中文的分词功能后，将该算法在实验环境中进行了测试。实验结果很不理想，虽然可以过滤 92.8%的垃圾邮件，但误报率达到了 6.3%。在垃圾邮件过滤系统中，误报一封正常的邮件所造成的损失远远超过漏报一封垃圾邮件所造成的损失，用户宁愿收到垃圾邮件也不会愿意自己的正常邮件被系

统误判并删除。因此这种高误报率的性能是不能被用户接受的。

通过以上的分析可见，在中文环境中任何一种算法都不能完全独立而又高效地过滤垃圾邮件。因此，本文提出一种综合的过滤方法，通过对这些现有算法的叠加和相互作用，来弥补各个单一算法自身的缺点和充分利用它们的优点，以提高邮件过滤系统的过滤能力。

4. 综合算法的设计

综合算法的目的是既保护正常邮件的收发，又可以高效地过滤垃圾邮件，并且使用户的额外操作减到最少，提高系统的自动学习能力。在综合算法中，系统对邮件进行多层过滤。其中白名单起到保护正常邮件的作用，规则算法在帮助系统过滤部分垃圾邮件的同时自动将结果传递给贝叶斯算法作为学习的资源，而贝叶斯算法是综合算法的核心，它在对规则算法过滤出的垃圾邮件、用户发出的正常邮件，以及用户的手动反馈信息等各种资源进行学习的基础上，将对最后一部分邮件进行计算和判断。综合算法中减去了黑名单功能，因为黑名单往往是一种惩罚性的措施，如果使用不当会产生误报，尤其是当封堵的对象是大型公共邮件服务器时会影响用户正常的接收邮件。

综合算法的总体结构图参见图 1，主要包括四个部分：邮件过滤功能、查看发出邮件功能、自学习功能和配置管理功能。图中“spam”表示垃圾邮件，“ham”表示正常邮件。

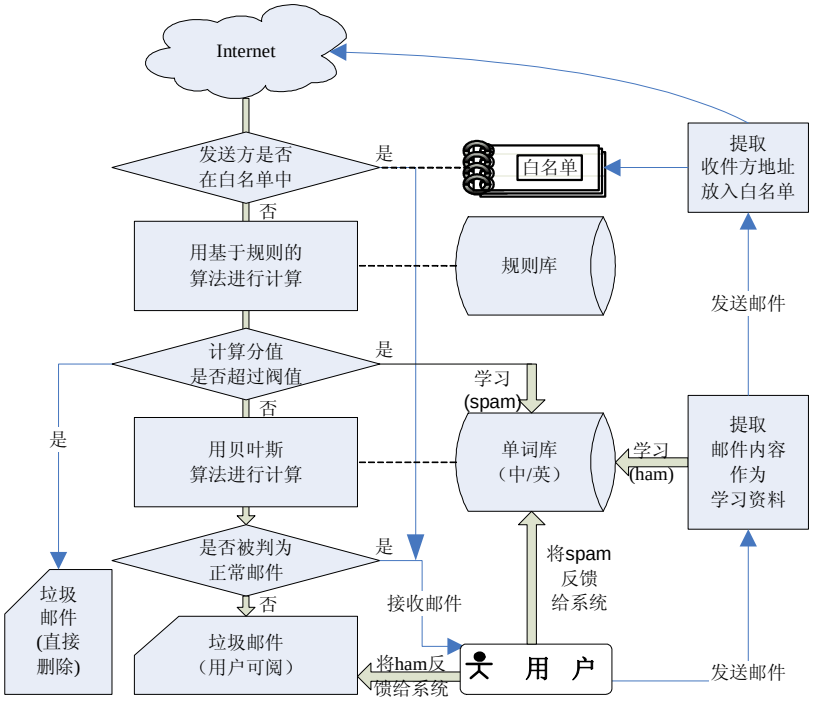

图 1 综合算法的总体结构图

Figure1. The structure of the integrated way

4.1 邮件过滤功能

邮件过滤功能对应于图 1 的左侧主线条。当一封新的邮件到达时，它需要经过白名单、基于规则的算法和贝叶斯算法的层层过滤。

对于一封新到达的邮件，系统首先查看邮件的具体发送方地址，如果是被列在白名单中，则该信件会被直接送到用户邮箱，否则继续用规则算法和贝叶斯算法进行判断。

综合算法中，基于规则的算法起到重要的辅助作用，系统要求该算法达到零误报率。为此除了提高阀值以外，还应根据实际情况调整规则的分值，使它更加适合于中文的实际环境。本文首先利用工具检查已知的正常邮件和垃圾邮件，通过计算各条规则在两类邮件中的匹配频率，从而判定中文环境中的邮件特征，并据此调整规则的分值。例如，有的规则在正常邮件中被大量匹配，说明它应被删除，或降低分值；而有的规则在垃圾邮件中大量匹配，但这

些垃圾邮件由于没有到达阀值而漏报，并且这些特征没有出现在正常邮件中，说明它在中文环境中需要提高分值。本文在实验中对任意选取的一个月的数据（垃圾邮件 52449 封，正常邮件 3862 封）做了一些统计，现将匹配频率最高的规则以及它们原先默认的分值、匹配频率分别列在表 1、表 2 中供参考：

规则名称	分值	匹配频率
MSG_ID_ADDED_BY_MTA_3	0.7	36.3%
HTML_MESSAGE	0.1	29.1%
MIME_HTML_ONLY	0.1	26.6%
MIME_HEADER_CTYPE_ONLY	1.9	23.9%
FORGED_MUA_OUTLOOK	3.5	22.4%

表 1 垃圾邮件中的规则匹配频率

Table1. Rule matching frequency in spam

规则名称	分值	匹配频率
USER_IN_WHITELIST	-100	38.3%
BASE64_ENC_TEXT	1.6	18.6%
NO_REAL_NAME	0.8	15.1%
HTML_MESSAGE	0.1	14.7%
MSG_ID_ADDED_BY_MTA_2	0.4	12.1%

表 2 正常邮件中的规则匹配频率

Table2. Rule matching frequency in ham

举例来说，表 2 中的 BASE64_ENC_TEXT 这条规则，是用来判定邮件是否以 Base64 编码方式传输，它的默认分值为 1.6。这是一个很高的分值，因为英文正常邮件中很少会出现 Base64 编码，只有英文的垃圾邮件常常以这种方式来躲避关键词的检查。但是在中文环境中，Base64 是一种很常见的编码方式，有 18.6%的正常邮件匹配到这条规则，因此这条规则的分值应当设为 0，即不作为判断依据。

由于系统中保证了规则算法的误报率为零，所以由它过滤出的垃圾邮件可以被直接删除，以减少用户的负担。

当新到达的邮件通过了规则算法的计算，最后将进入贝叶斯算法的判断。贝叶斯算法在自学习模块的学习和知识积累的基础上，对这封邮件的全文进行查看和计算，以帮助用户做出判断。虽然综合算法将通过各种保护措施使误判的可能性减少到最低，但由于系统不能保证绝对的正确，因此由贝叶斯算法过滤出的这少部分垃圾邮件会被保存一段时间供用户适时地检查，这同时也是为了在用户怀疑是否有正常邮件被误判时可以有据可查。

4.2 查看发出邮件功能

由图 1 可见，系统对用户发出的邮件都将进行处理，提取接收方地址列入白名单中，并将邮件的内容作为单词库中正常词的主要来源。

本文提出从用户发出的邮件中自动提取白名单。按照常理，用户发出邮件的对象肯定是友好的，并且发出邮件的地址也一定是真实的。不仅如此，用户对于一封新到达的正常邮件通常会给予回复，而一旦回复，这封新邮件的发送方地址自然就会作为发出地址被记录下来。这样做的好处有两点：1)白名单信息是在用户正常使用邮件的过程中被自动获取的，用户无需额外的操作；2)白名单具有很高的可信度，它的正确性不会因为过滤系统的误判而降低。

同理，用户发出的邮件还将作为贝叶斯算法进行正常邮件学习的主要来源。用户发出的邮件内容通常与收到的正常邮件的内容相似，使用的语言习惯也是相通的。并且，用户在回复一封正常邮件时往往会附带上原信件的内容，这些内容对贝叶斯算法来说将是宝贵的学习资源。

4.3 自学习功能

针对贝叶斯算法过于依赖用户反馈的缺点，本文综合了各种学习渠道，增加了系统自动获取知识的能力，从而能够不断地更新单词库的内容，保持系统的效率。

图 1 中的四条粗箭头表示贝叶斯算法的主要学习过程。学习资源中有一类是由系统自动获取，另一类来自于用户的手动反馈。第一类的资料来源是综合算法自动学习的核心体现，它从其它方法的结论中得到辅助信息，包括由规则算法过滤出的垃圾邮件将被自动传给贝叶斯算法进行垃圾词的学习，以及用户发出的邮件内容将自动作为正常词学习的重要来源。由

于这一类资源的内容是丰富和动态变化的，系统可以由这一类资源不断地获取和更新知识，由此系统对第二类资源的依赖性将明显降低，也就是用户即使对系统不做出任何反馈，系统也可以从第一类资源中自动学习新知识，并发挥良好作用。

但是在实际运用中，系统仍然鼓励用户做出及时反馈，例如将漏报的垃圾邮件转发给系统，或者定期地检查用户可阅读的垃圾邮件。这样，用户通过信息反馈可以帮助系统更加准确的判断邮件性质，使系统的总体性能更加优越。

4.4 配置管理功能

普通用户可以通过配置选项打开和关闭过滤器的使用，也可以手动调整白名单。对于后者，普通用户通常无需做改动，因为系统中的白名单是可以自动搜集和维护的。管理员可以查看和管理所有用户的配置内容。

综合算法通过这四个部分的相互作用，从用户发出的邮件中动态获取白名单，并且通过将规则算法过滤出的垃圾邮件和用户发出的正常邮件作为学习资源传递给贝叶斯算法进行自学习，从而自动保持贝叶斯算法的效率。用户的反馈可以使系统的判断更加精确。

5. 综合算法的效率分析

综合算法通过各种过滤方法的相互结合，实现了系统的自动学习和动态更新，并使中文环境中的垃圾邮件过滤效率有较大提高。

由表 2 中 USER_IN_WHITELIST 这条规则的匹配频率可知，当白名单设为静态列表时，它可以保护 38.3%的正常邮件。本文对收到的正常邮件作统计，其中有 83.8%的邮件来自于经常联系的好友，而在陌生地址中又有 7.6%是回复信件，由此可以推算，利用综合算法仅从“查看发出邮件功能”中动态得到的白名单就可以保护 91.4%的正常邮件。

贝叶斯算法是提高系统正确性的关键，系统通过自动获取辅助学习资源的方式弥补了该算法过于依赖用户反馈信息的缺点。实验中，经过分值调整的规则算法可以在零误报率的前提下过滤 71.4%的垃圾邮件，它们在被删除前都将经过贝叶斯算法的学习。另一方面，从正常邮件的统计数据可知，26.5%的正常邮件会被用户回复，正常邮件中有 22.1%是回复给用户的信件，其中的原件内容都会在发出邮件中被自动获取。并且，通过观察发现，其余的不被回复的邮件往往是对贝叶斯学习帮助不大的内容，例如群发的娱乐信件、附件资料、系统退信等。所以用户没有增加任何的额外操作，就可以帮助系统取到最有价值的学习内容，体现了综合算法中自动学习的优势。

最后，本文针对同样的邮件样板（任意挑选的正常邮件和垃圾邮件各 1000 封），分别采用三种单一的过滤算法和本文的综合算法进行比较。比较结果如表 3 所示，其中查全率表示垃圾邮件样板中，被正确识别为垃圾邮件的比例；误报率表示正常邮件样板中，被误判为垃圾邮件的比例。

	黑名单 (IP 级地址列表)	规则算法(对各分值和总阀值已作调整)	贝叶斯算法(对中文采用双字分词)	综合算法(无手动反馈)
查全率	5.4%	71.4%	92.8%	96.6%
误报率	0	0	6.3%	<1.6%

表 3 三种单一算法与综合算法的比较

Table3. Comparison between the three single algorithms and the integrated way

表 3 中，单一的贝叶斯算法是各用前 500 封邮件先进行学习，再用后 500 封邮件进行测试得到的结果。而综合算法的结果是在没有用户手动反馈信息的前提下得到的，如果实际使用中用户能够给予反馈，其效率将会更高。综合算法的查全率和误报率可以根据样板邮件的内容和算法描述的功能计算出来：查全率采用垃圾邮件样板进行计算，首先用规则算法对样板垃圾邮件进行过滤，然后将规则算法过滤出的垃圾邮件作为学习资料交给贝叶斯算法学习，最后用贝叶斯算法对剩下的邮件进行过滤，总共有 96.6%的垃圾邮件被识别；误报率采

用正常邮件样板进行计算，首先滤除从友好地址发来的邮件，然后将主题为“回复”的邮件滤除，因为回复邮件的发送方地址在用户发出邮件时就应该被记录下来。再将主题为“回复”邮件中的原件内容，即用户先前发出的邮件内容，作为学习资料交给贝叶斯算法学习，并对剩余邮件进行判断，此时误判率已降为 1.6%。误判的邮件中除了期刊邮件、网站退信等内容外，有一类邮件占了很大比例，即学生提交的作业。如果任选其中一封作业邮件作为正常邮件再交给贝叶斯算法学习，那么这一类邮件都不再会被误判，总体的误判率降到 0.9%。由此可见，用户的反馈仍然很重要。另外，如果能够获取用户发出的所有邮件作为贝叶斯算法的学习资料，总的误判率应该会更低。

贝叶斯算法作为综合算法的核心，是提高总体效率的关键所在。为了使系统达到更高的效率，使误判率接近于零，还应该考虑对贝叶斯算法做三个方面的改进。首先是提高中文的分词技术。本文实验中使用的双字分词法是造成误报的原因之一。单词库中有些单词是无意义的，并且均是两个字的单词，它们不仅对正确的判断不起作用，而且还会干扰概率计算。所以，算法改进的第一步是改变中文的分词技术，如采用基于词典的分词算法等。关于中文的处理是一个专门的研究领域，目前 SIGHAN 等组织正致力于这方面的研究。

第二是需要区分单词库中的中英文单词。尽管实验环境中以中文邮件为主，但英文邮件仍然需要谨慎处理，且英文邮件的比例有逐渐增大的趋势。在实验中，有很多英文邮件被误报为垃圾邮件。这是因为现在的单词库中以中文单词为主，只有很少量的英文单词来自于信件头，所以一封英文邮件只有信件头中的单词能够在计算中起到作用，这样很容易被判为垃圾邮件。在改进的贝叶斯算法中，应将中文和英文的单词库完全区分开来，并对中英文邮件采用各自的单词库来计算概率。

第三是建立单词库中的用户组。贝叶斯算法具有个性化的特征，很多基于该算法的工具都是安装在客户端。为了减少用户的操作和方便管理，本文的综合算法是应用在服务器端，但在使用贝叶斯算法时，必须保持它的个性化，主要体现在单词库上。根据实验环境的用户分类特征，可以将用户分为多个组，按组建立单词库。极端情况下，可以单个用户分为一个组。

本文中贝叶斯算法的检测率没有 Paul Graham[3]的工具的效率高是因为后者使用于英文环境，而本文所针对的是中文环境，并且没有使用手工的词库。如果能够对贝叶斯算法进行以上三方面的改进，尤其是中文单词的自动划分方法，系统整体的检测准确率可以进一步提高。

综合算法的优越性还体现在系统使用和管理的方便程度上。无论是黑白名单算法，还是基于规则的算法都需要用户不断的更新和维护操作，而单一的贝叶斯算法也必须依赖于用户的手动反馈，因此任何一种单一的算法都需要增加用户大量的额外操作。而综合算法可以进行自动的学习和工作，不需要用户甚至管理员过多的干预。用户的反馈不再是必须的，而只是受到鼓励的行为，当用户愿意反馈一些信息时，系统会在效率上给用户予回报。

6. 结束语

本文针对目前日益严重的垃圾邮件问题，提出了一个适合中文邮件环境的垃圾邮件综合过滤方法。它以动态维护的白名单来保护正常邮件的收取，以规则算法作为辅助过滤手段，采用用户发出的邮件作为重要的学习资料来源，并将贝叶斯算法作为核心来提高邮件过滤的准确度，从而使整个系统可以在保护正常邮件的基础上，自动地帮助用户过滤垃圾邮件。

该综合过滤方法是根据真实的中文邮件环境特征，在对原始邮件数据进行实验分析和数据统计的基础上提出的，并采用实际数据进行了效率分析，具备较高的可信度和实用性。但该综合方法在实际应用中还需完成贝叶斯算法本身的改进，从而使系统的准确率更高。

本文的综合过滤方法仅针对内容上不被请求的垃圾邮件而言，对于蠕虫病毒引起的垃圾邮件不能很好地过滤。蠕虫邮件以传播自带的病毒附件为目的，很少带有大量文本内容，其

发送方地址多为伪造地址，并且具有很强的周期性，因此对蠕虫邮件和内容上的垃圾邮件的过滤方法是完全不同的。

参考文献

[1] Paul Graham. “A Plan for Spam.” Aug 2002. http://www.paulgraham.com/spam.html

[2] Paul Graham. “Stopping Spam.” Aug 2003. http://www.paulgraham.com/stopspam.html

[3] Paul Graham. “Better Bayesian Filter.” Jan 2003. http://www.paulgraham.com/better.html

[4] Patrick Pantel, Dekang Lin. “SpamCop-- A Spam Classification & Organization Program.” Proceedings of AAAI-98 Workshop on Learning for Text Categorization

[5] Mehran Sahami, Susan Dumais, David Heckerman and Eric Horvitz. “A Bayesian Approach to Filtering Junk E-Mail." Proceedings of AAAI-98 Workshop on Learning for Text Categorization