基于 Agent 的邮件过滤与个性化分类系统设计

刘 毅,张月琳
(东南大学 计算机科学与工程学院,江苏 南京 210096)

摘 要:随着电子邮件的广泛使用,垃圾邮件的危害日益增大,用户的个性化需求也日趋强烈。文中提出了一个基于 Agent 的邮件过滤与个性化分类系统,既能过滤垃圾邮件,又可以根据用户的个性化需求对正常邮件进行自动分类。垃圾邮件过滤采用了朴素贝叶斯方法,邮件的个性分类采用了最小风险贝叶斯方法。邮件个性化分类有效地利用了邮件过滤的输出,提高了系统运行的效率。本系统还可以接收用户的反馈并传递给对应的 Agent,从而改进分类算法,不断地微调分类系统。

关键词:Agent;邮件分类;特征选择;朴素贝叶斯方法;最小风险贝叶斯方法
中图分类号:TP393.098 **文献标识码**:A **文章编号**:1673-629X(2009)02-0066-04

Design of a Mail Filter and Personalized Classification System Based on Agent

LIU Yi, ZHANG Yue-lin
(School of Computer Science & Engineering, Southeast University, Nanjing 210096, China)

Abstract: With the wide application of E-mail, the spam does more and more hurt, and the users want more personalized services. A mail filter and personalized classification system based on Agent is provided. This system can not only filter the spam, but also classify the normal mail automatically according to the users' personalized requirements. The spam filter uses naive Bayes method, and the mail personalized classification uses minimum risk Bayes method. The mail personalized classification makes good use of the output of the spam filter, as a result improves the system's operating efficiency. This system can get users' feedback and send to the corresponding agent, in order to improve the classification algorithm and make delicate adjustment to the classification system.

Key words: agent; mail classification; feature selection; naive Bayes method; minimum risk Bayes method

0 引 言

随着互联网的高速发展,电子邮件以其方便、快捷的特点成为人们在日常交流中的重要工具。电子邮件的广泛使用带来的是垃圾邮件的泛滥。垃圾邮件不但浪费收信者的宝贵时间,而且消耗了大量的网络资源,更严重威胁到网络安全,成为必须治理的问题。此外,在对正常邮件的处理上,用户往往会根据需要、喜好建立自己的个性化文件夹,并在收到邮件后对邮件进行归档,但对大量的邮件进行手工归档是一件繁琐的工作。文中设计了一个基于 Agent 的邮件过滤与个性化分类系统,对收到的邮件进行过滤区分出垃圾邮件和正常邮件,然后根据用户建立的类别把正常邮件分类存放,满足个性化的分类管理需求。

收稿日期:2008-05-08
作者简介:刘 毅(1982-),男,江苏南通人,硕士研究生,主要研究方向为计算机系统结构;张月琳,教授,主要研究方向为计算机系统结构。

1 Agent 技术

Agent 是在人工智能领域发展起来的概念,目前还没有一个统一的定义。一般来说,Agent 是一个运行于动态环境的自治体,并具有以下特征[1,2]:

(1)自主性:Agent 具有属于其自身的计算资源和局部于自身的行为控制机制,能够在没有外界直接操纵的情况下,根据其内部状态和感知到的环境信息,决定和控制自身的行为。

(2)反应性:Agent 能够感知所处的环境,并对所处的环境的变化适时地做出反应。

(3)主动性:Agent 不仅能对其所处的环境做出反应,还能够遵循承诺采取主动行动,表现出面向目标的行为。

(4)面向目标的能力:为了更好地实现目标,Agent 能够将复杂的任务分解为多个子任务,并决定子任务的执行顺序和方法。

(5)交互性:Agent 能够通过某种 Agent 通信语言

与其他 Agent 或人进行交互。

2　系统的结构与组成

2.1　系统的结构

图 1 是基于 Agent 的邮件过滤与个性化分类系统的结构。本系统从功能上可以划分为两大模块:邮件过滤模块与个性化分类模块。邮件过滤模块以用户收到的正常邮件作为输入,输出为分成两类的邮件:垃圾邮件和正常邮件。个性化分类模块以过滤模块输出的合法邮件为输入,输出为归档到用户自定义分类的邮件。本系统具有自适应性,能够接受用户的反馈,并通过学习改进分类算法,使得系统能更好地为用户提供服务。

2.2　系统的组成

本系统由六个 Agent 组成:

(1)文本分词 Agent。

文本分词 Agent 是一个基本功能 Agent,将输入的原始邮件转换为词条集。此 Agent 将负责的任务分成两个子任务:邮件预处理,文本分词。

(1.1)邮件预处理。

电子邮件是一种半结构化的文本文件,由邮件头和正文组成。邮件头包含邮件传递过程中经历的 MTA、发送者与接收者、主题、日期等信息。预处理分两步:第一步去掉无用的结构信息,只保留主题和正文的文本;第二步从文本中删除那些出现频率很高但与邮件特征无关的词(如连接词、语气助词等),起到初步降低噪音的效果。

(1.2)文本分词。

文本分词是在中文词典的支持下,把一封邮件的文本切分成有意义的中文词条序列,构成词条集。

为了提高文本分词时查询词典的速度,中文词典的存储结构由 n 张哈希表构成(n 是词条最大长度),分别存储 n 字词、$n-1$ 字词、…… 双字词和单字词。词条在哈希表中的位置由词条的哈希码决定。设词条 X

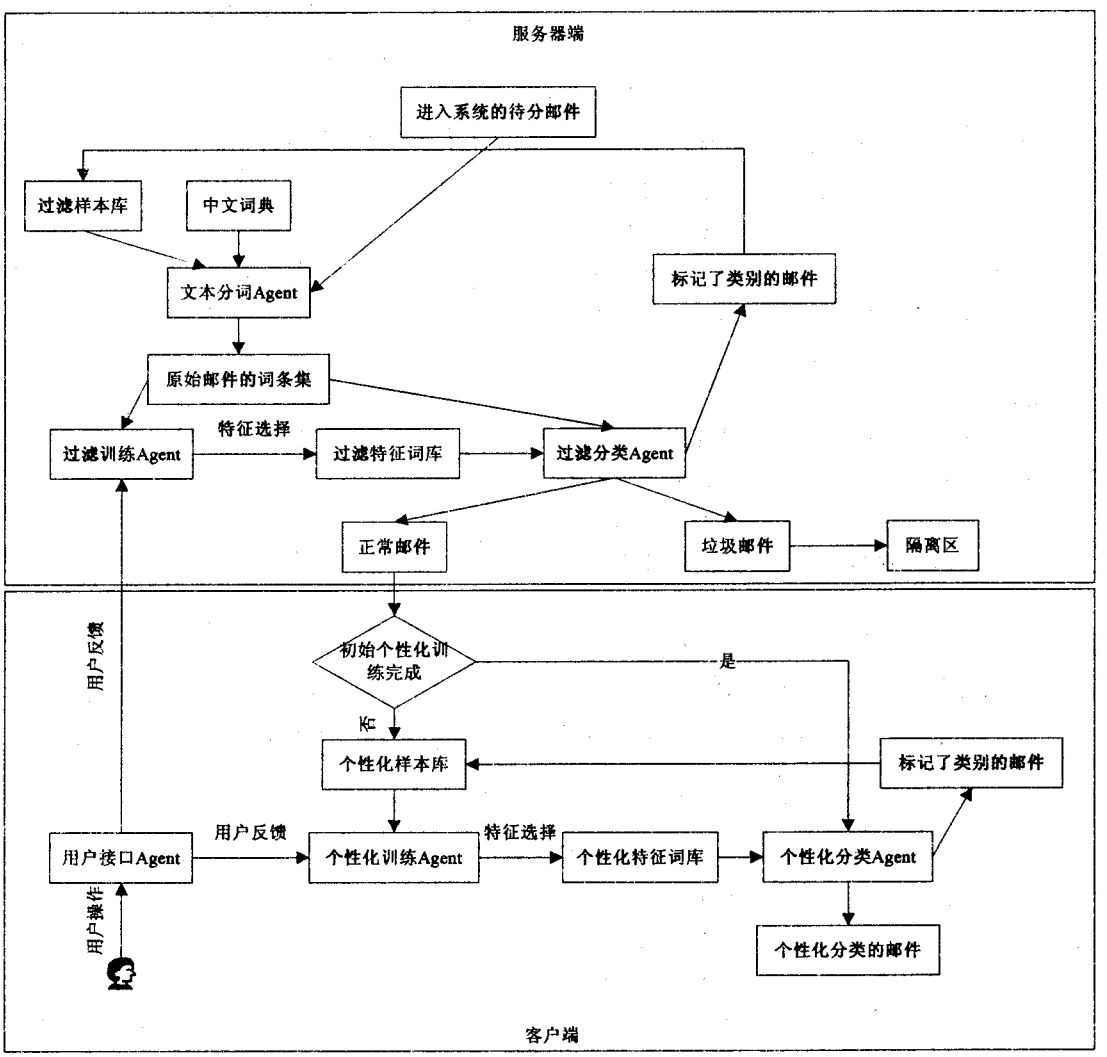

图 1　系统结构图

的长度为 n，由字符序列 $X[0],X[1],\cdots X[n-1]$ 组成，则该词条的哈希码计算公式为：$\text{HashCode}(X) = X[0]^n + X[1]^{(n-1)} + \cdots + X[n-1]$，式中的 $X[i]$ 是字符 $X[i]$ 的 Unicode 编码值[3]。

本系统使用的文本分词方法是逆向最大匹配法，流程如图 2 所示。

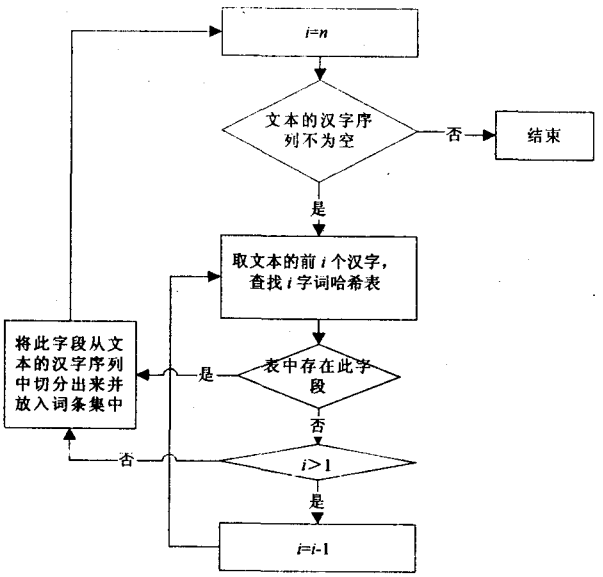

图 2 逆向最大匹配法流程图

从图 2 流程可以看出，词条最大长度 n 的选择对文本分词的执行是有影响的。若最大长度过小会造成长词条的人为截断，若最大长度过大对于较短的词条来说会进行多次无效的切分。本系统选择词条最大长度为四，兼顾了文本分词的正确率与效率。

(2)过滤训练 Agent。

过滤训练 Agent 的目标是生成过滤特征词库。它将过滤样本库中的邮件交由文本分词 Agent 处理，生成并合并所有样本的词条集，然后从合并后的词条集中选出对分类邮件最有效的特征词条从而构成特征词库。

特征选择是邮件分类的基础。由于从样本邮件生成的原始词条集包含的数据维数过大，将之直接用于构建决策系统是不可能的，必须通过特征选择去除出现频率过低或对分类贡献不显著的词条，达到降低维数的目的。不同的特征选择算法选出的特征词库差异很大，对邮件分类的准确率也会产生很大的影响。本系统选用信息增益法作为特征选择的方法[4]。

信息增益法的基本原理是定义一个参数来衡量词条对分类的贡献，对所有的词条计算出这个参数后进行排序，选出贡献度大的若干词条构成特征词库。样本邮件的类别是事先指定的，假设有 m 类，定义为 $c_1, c_2,\cdots,c_m$。t 为某个词条，$\text{IG}(t)$ 代表 t 的信息增益，即 t 对分类的贡献度。对于过滤训练样本来说，$m = 2$(垃圾邮件和正常邮件)。

$\text{IG}(t)$ 的计算公式如下：

$$\text{IG}(t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t) + P(\bar{t})\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t})$$

其中 $P(c_i)$ 表示 c_i 类邮件在样本集中出现的概率，$P(t)$ 表示包含词条 t 的邮件在样本集中出现的概率，$P(c_i \mid t)$ 表示邮件包含词条 t 时属于 c_i 类的条件概率，$P(\bar{t})$ 表示不包含词条 t 的邮件在样本集中出现的概率，$P(c_i \mid \bar{t})$ 表示邮件不包含词条 t 时属于类 c_i 的概率。

计算所有词条的信息增益，按从大到小排列，设置一个阈值 d，将所有信息增益值大于 d 的词条加入特征词库。

信息增益法只考虑词条出现与否，忽略了词条的出现频率的信息。由于训练样本比较充足，即使在计算中不加入词频因素也能很好地估计词条对分类的贡献度，过多地考虑词频反而有可能误导词条对分类的影响。

(3)过滤分类 Agent。

过滤分类 Agent 利用文本分词 Agent 得到待分类邮件的词条集，然后在过滤特征词库的支持下，用最小风险贝叶斯方法判断邮件的类别——正常邮件或垃圾邮件。

(3.1)朴素贝叶斯方法[5,6]。

贝叶斯方法的基本原理是：一个事件将来发生的概率可以从它以前发生的概率推断得到。假设类别集为 $C = \{c_1, c_2, \cdots, c_m\}$，给定一封邮件 d，其词条集为 $\{t_1, t_2, \cdots, t_n\}$。$d$ 属于 c_i 类的概率为 $P(c_i \mid d) = \dfrac{P(c_i)P(d \mid c_i)}{P(d)}$，其中，$P(c_i)$ 是 c_i 类邮件出现的概率，$P(d \mid c_i)$ 是 c_i 类邮件中出现邮件 d 的概率，$P(d)$ 是邮件 d 出现的概率。

由于邮件 d 的各词条的取值相互关联，计算 $P(d \mid c_i)$ 过于复杂，因此采取朴素贝叶斯方法：假设各词条取值相互独立，此时 $P(d \mid c_i) = \prod_{k=1}^{n} P(t_k \mid c_i)$。

使 $P(c_i \mid d)$ 取最大值的类 c_i 为邮件 d 所属类别，即

$$c(d) = \arg_{c_i \in C}\max \frac{P(c_i)\prod_{k=1}^{n} P(t_k \mid c_i)}{P(d)}$$

由于 $P(d)$ 对于 (c_i) 是常数，上式简化为 $c(d) = \arg_{c_i \in C}\max P(c_i)\prod_{k=1}^{n} P(t_k \mid c_i)$，其中 $P(c_i) = \dfrac{N_i}{N}$，N_i 是

训练样本中 c_i 类邮件的个数,N 是训练样本的总数。

在计算 $P(d \mid c_i)$ 时,如果某个词条 t_k 在 c_i 类邮件中不出现,此时 $P(t_k \mid c_i) = 0$,从而导致乘积 $\prod_{k=1}^{n} P(t_k \mid c_i)$ 为零,即 $P(d \mid c_i) = 0$。为了避免这种情况的出现,用下式估计 $P(t_k \mid c_i)$:

$$P(t_k \mid c_i) = \frac{n_{ki} + 1}{n_i + \mid V \mid}$$

其中,n_{ki} 为 c_i 类邮件中词条 t_k 出现的次数,n_i 为 c_i 类邮件中特征词条的总数,$\mid V \mid$ 为特征词库的大小。

(3.2) 最小风险贝叶斯方法。

在实际的分类过程中,做出决策总要承担一定的风险,而不同的决策所承担的风险大小是不一样的,比如将正常邮件当作垃圾邮件过滤与漏过一封垃圾邮件相比,前者造成的损失要大得多。最小风险贝叶斯方法正是考虑了风险因素而提出的一种决策规则[7]。

最小风险贝叶斯方法引进了损失因子 $\lambda(a_i, c_j)$;$i = 1,2,\cdots,m$;$j = 1,2,\cdots,m$。$\lambda(a_i, c_j)$ 表示将 c_j 类的邮件判断为 a_i 类所造成的损失。

最小风险贝叶斯方法如下:

① 对每个类别 c_i,计算 $P(c_i \mid d)$;

② 对每个 a_i,计算将邮件 d 判断为 a_i 类需要承担的风险度:

$$R(a_i \mid d) = \sum_{j=1}^{m} \lambda(a_i, c_j) P(c_j \mid d)$$

③ 选择风险度最小的决策 a_k,将之作为邮件 d 的类别。

标记了类别的邮件作为新的训练样本被送入过滤样本库。判断为垃圾邮件的文档并不直接删除,而是暂时保存在服务器上的隔离区中,当隔离区满后根据日期删除最早的邮件。用户可以从隔离区找回被误判的正常邮件,或是将漏判的垃圾邮件移到隔离区中,这些转移邮件的操作被用户接口 Agent 记录,作为用户的反馈传递给过滤训练 Agent,过滤训练 Agent 接收到反馈后更新样本库中对应邮件的类别。当过滤样本库增长到一定规模后,过滤训练 Agent 开始新一轮的训练。

以上三个 Agent 运行于服务器端,为所有邮件用户提供统一的邮件过滤服务。

(4)个性化训练 Agent。

个性化训练 Agent 与过滤训练 Agent 的目标类似,都是生成用于邮件分类的特征词库,区别在于:个性化样本库并非事先给定,而是由用户的正常邮件生成,因此训练样本的规模是随时间逐渐增长的。当训练样本很少的时候,估计词条的信息增益会有很大的误差,导致分类的准确率极低。为了保证训练在样本充足的前提下进行,系统设置了样本规模的阈值,只有当样本数目超过阈值后才会开始初始个性化训练。

在初始个性化训练前的样本积累阶段,每一封通过过滤的正常邮件被送入个性化样本库(邮件在过滤阶段已经被分词,因此直接以词条集的形式进入样本库),标记为默认类别,并发送到用户的默认收件箱。当用户将某封邮件从默认收件箱转移到个性化文件夹时,个性化训练 Agent 从用户接口 Agent 接收到用户反馈,将这封邮件的对应样本标记为个性化文件夹的类别。举个例子:张三建立了一个名为"工作"的文件夹。当他收到邮件 d,并将 d 转移到"工作"文件夹时,d 在样本库中的词条集类别就被标记为"工作"了。由此可见,个性化训练 Agent 的样本类别划分依赖用户的手工归档,类别集 C 为默认类别和用户所有的个性化文件夹类别。

样本积累完毕后,初始个性化训练开始进行,流程与过滤训练一致,生成个性化特征词库。

(5)个性化分类 Agent。

当初始个性化训练完成后,通过过滤的正常邮件被送入个性化分类 Agent。个性化分类 Agent 在个性化特征词库的支持下,用朴素贝叶斯方法判断邮件的类别,将邮件发送至对应类别的个性化文件夹。分类算法不用最小风险贝叶斯方法的原因是:类别是由用户指定的,判断错误造成的损失因人而异,没有一个统一的标准。

标记了类别的邮件作为新的训练样本被送入个性化样本库。当用户在个性化文件夹之间转移邮件时,他的操作被用户接口 Agent 记录,作为用户反馈传递给个性化训练 Agent,个性化训练 Agent 接收到反馈后更新样本库中对应邮件的类别。当个性化样本库增长到一定规模后,个性化训练 Agent 开始新一轮的训练。

(6)用户接口 Agent。

用户接口 Agent 负责接收用户的反馈。它记录用户执行的各种影响邮件分类的动作,如从隔离区中找回邮件,将邮件移至隔离区,将邮件从一个文件夹移至另一个文件夹等,并将用户的反馈传递给训练 Agent,训练 Agent 得到反馈后完善样本库,改进分类算法,分类系统也随之不断地进行微调。

以上三个 Agent 运行于客户端,为每个用户提供个性化分类服务。

3　Agent 在邮件过滤与个性化分类中的优势

Agent 自身的特点决定了它很适合用于邮件过滤

(下转第 73 页)

的数据基本一致，那么该数据的可靠性就越大。通过此法更新数据库中的信息，同时更新下载表中的有效性字段值。

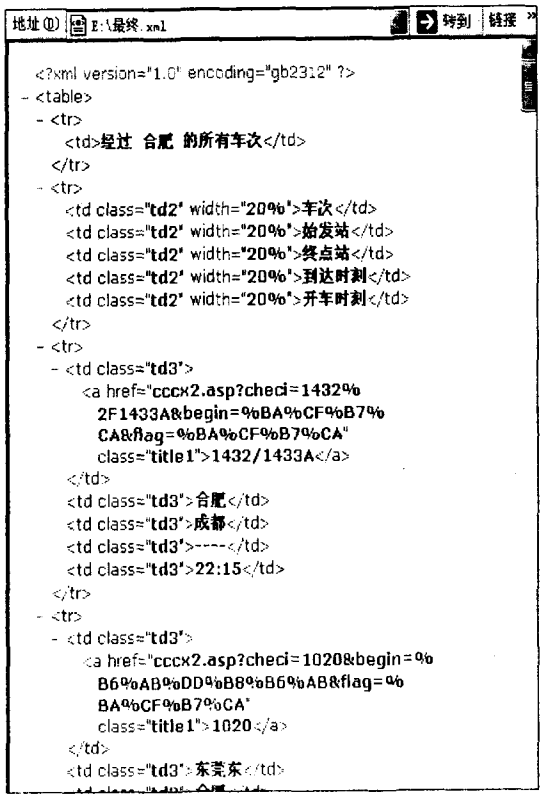

图 6　后处理后的.XML 文件

2.2.4　前台检索

利用 Visual Studio 2005 集成开发环境和 C# 语言，可以很轻松地开发 Asp. net 网站。我们编写建立自己的网页，利用下载到数据库中的信息，可以提供给用户多方位和快捷的服务。需要指出的是，前台页面的数据检索是基于信息表，但由于其中的数据不具有一致性，需要经过重组织方能使用，方能提供给用户完善的信息。

3　结束语

作为 Web 数据挖掘的一次实践性研究，设计的系统还不够完善，在实际的设计中还存在很多问题，比如：因为是从网上下载数据，而提供网页的服务商会更新网页，从而引起网址格式的变动，这必然使大量下载成为难题；各个网页提供的数据结构不一，信息也不一致(有的信息不完整)，这对提取工作提出挑战；同样，存入数据库的信息的不一致性，也给检索算法的设计带来麻烦。这些都是亟待解决的问题。

总的来说，Web 数据挖掘目前还处于发展时期，其技术还不成熟，但由于其广泛的应用价值，必然会让更多的人来参与研究，其路会越走越远。

参考文献：

[1] 苏新宁，杨建林，邓三鸿，等. 数据挖掘理论与技术[M]. 北京：科学技术文献出版社，2003.

[2] 恽　爽，韩立新，董　浚，等. KDW 综述：基于 Web 的数据挖掘[J]. 计算机工程，2003(1)：284－286.

[3] Linoff G S，Berry M J A. Web 数据挖掘：将客户数据转化为客户价值[M]. 沈钧毅，等译. 北京：电子工业出版社，2004.

[4] 李雄飞，李　军. 数据挖掘与知识发现[M]. 北京：高等教育出版社，2003.

[5] 周琪锋. 基于 Web 的数据挖掘技术的研究[J]. 电脑知识与技术，2007(1)：97－103.

(上接第 69 页)

与个性化分类。

Agent 的自主性使得 Agent 能够在没有管理员指导的情况下由自身的决策机制决定采取何种行动。

Agent 的反应性使得 Agent 能够学习用户的反馈，不断调整和改进分类算法，完善分类系统。Agent 的社会性使得系统的各组成 Agent 能够方便地相互交流，更好地为用户服务。

4　结束语

文中提出了一种基于 Agent 的邮件过滤与个性化分类系统，不但可以过滤垃圾邮件，还可以对正常邮件进行自动分类，有效地减少了用户手工归档的工作量。目前，该系统的具体实现还在实验之中，有关 Agent 与 Agent 之间的高效通信等问题还有待进一步研究。

参考文献：

[1] 徐俊萍，翟玉庆. 基于 Agent 的个性化信息服务技术的研究[J]. 计算机工程与科学，2002，24(3)：74－76.

[2] 张　克，邵长胜，强文义. 基于面向 Agent 技术的任务规划系统研究[J]. 高技术通讯，2002(5)：82－86.

[3] 王新梅，芦　苇，尹朝庆，等. 基于文本挖掘的邮件分类与过滤[J]. 计算机工程与应用，2006(2)：135－137.

[4] 李志博，余正红，尹朝庆，等. 邮件服务智能代理的研究[J]. 计算机工程与设计，2007，28(3)：683－686.

[5] Zhang PHaiyi，Li PDi. Naive Bayes Text Classifier[C]//Proceedings of the 2007 IEEE International Conference on Granular Computing. USA：[s. n.]，2007.

[6] 成宝国，冯宏伟. 一个基于 Naive Bayesian 垃圾邮件过滤器的改进[J]. 计算机技术与发展，2006，16(2)：98－99.

[7] 王　宁，张建忠，何　云，等. 基于改进贝叶斯模型的中文邮件分类算法[J]. 计算机工程与应用，2006(31)：97－100.