计算机科学 2002Vol. 29№. 8

基于最小风险的贝叶斯邮件过滤算法*)

Mail Filtering Based on the Risk Minimization Bayes

石霞军　林亚平　陈治平

(湖南大学计算机与通信学院　长沙 410082)

Abstract　A new algorithm of filtering junk-mail based on the risk minimization Bayes is proposed in this paper, which can reduce the error rate of misclassifying legal mail. The experiment results show that the algorithm has better performance.

Keywords　Feature extraction, Risk minimization Bayes, Mail filtering

1. 引言

随着因特网的迅猛增长,电子邮件作为最快捷、最经济的通信方式,也得到飞速发展。但是,许多销售广告、快速致富等垃圾邮件也在网络中传送,这些垃圾邮件不仅占据邮件服务器的大量存储空间,同时也要用户花费大量时间来处理这些垃圾。尽管一些商业化产品允许用户人工建立垃圾邮件的过滤规则,但是系统要求用户有丰富经验和花去许多时间,而且,由于垃圾邮件在不断改变,用户必须经常调整这些规则,这需要花大量时间。因此,研究邮件自动过滤方法具有重要意义。邮件自动过滤方法研究主要有基于规则和基于概率[1]两种,后者已成为一种主要研究趋势。

文[5]通过动态学习规则将邮件进行分类,这种方法对基于邮件内容进行一般分类时具有一定作用,但在邮件过滤过程中没有考虑到垃圾邮件本身具有与合法邮件不一样的特性,而且文中提出的 RIPPER 算法需要消耗大量的时间。

Sahami[2]利用贝叶斯(Bayes)算法对邮件进行分类,并结合手工构造的短语以及具有垃圾邮件特性的单词来提高过滤垃圾邮件的精确度。但作者为减少合法邮件因被误判为垃圾邮件,将评判为垃圾邮件的概率门槛值提高到了 99.9%。这种过高的门槛值造成大部分垃圾邮件被误判为合法邮件的现象。

Rennie[4]在基于贝叶斯算法的基础上建立了一个用于邮件过滤的机器学习应用系统 ifile,利用贝叶斯算法对邮件进行分类,这种方法分类速度比较快,不需要用户太多干预,错误分类的邮件可以通过用户对该邮件的不同类间的转移,并利用贝叶斯算法重新训练,可得到动态调整。但是由于错误分类的现象比较多,用户仍需要检查垃圾邮件中是否存在错误分类的合法邮件而要花费大量时间。

本文在朴素贝叶斯算法的基础上,基于最小风险贝叶斯方法提出一种新的邮件过滤算法,在尽量不使合法邮件错判为垃圾邮件的同时,尽可能地提高分类的精确度。实验表明,这种基于最小风险贝叶斯邮件过滤算法具有较好性能。

2. 朴素贝叶斯(Naive Bayes)算法

朴素贝叶斯算法是在一般贝叶斯算法的基础上通过假定各因素之间不存在任何联系,即完全独立而得到的一种简化贝叶斯算法。这种算法在文本分类中得到非常广泛的应用。根据贝叶斯概率公式,对于给定的向量 $d(\omega_1,\omega_2,\cdots,\omega_n)$ 属于第 $C_k(k=1,2,\cdots,m)$ 类的概率为:

$$P(C_k|d)=P(C_k)\times P(d|C_k)/P(d) \qquad (1)$$

其中:$P(d)=\sum_{k'=1}^{m}P(d|C_{k'})\times P(C_{k'})$

由式(1)可知,要判断一个待识别邮件的类别,可以通过计算 $P(C_k|d)$ 概率来完成,它表示出该文档中出现的单词与向量空间模型中特征项的匹配情况,而决定该文档属于第 C_k 类的概率。我们可通过先验概率 $P(C_k)$ 和条件概率 $P(d|C_k)$ 来得到后验概率 $P(C_k|d)$。

假定 w_j 表示第 j 个特征项,基于文档中单词出现的概率相对独立的假设,有:

*)湖南省自然科学基金资助项目(99JJY20060)。石霞军　硕士研究生,研究方向为计算机网络。林亚平　教授,博士生导师,主要研究方向为网络计算和机器学习。陈治平　博士研究生,研究方向为机器学习。

～55

2 Bahk S, Zarki M E. Congestion control based dynamic routing in ATM networks. Comput. Commun., 1994, 17(12): 826～842

3 Tao yang, et al. Blocking Infection of Network and It's Diagnosing and Analyzing. 2000 Intl. Conf. on Communication Technology Proceedings, IEEE Press (IEEE catalog number: 00EX420)& Publishing House of Electronic Industry, Beijing: China, 2000. 296～301

4 Tao yang. The Analysis of the Congesting Control Between the Networks. In: Proc. ISTIST'96 (1996 Intl. Symposium on New Transmission & Switching Technologies), Kun-ming: China, 1996. 318～321

5 Tao Yang, et al. The Analysis of Buffer Module Stream Based on Cross-switch Structure. In: Proc. ISTN'97 (1997 International Seminar on Teletraffic and Network), 1997 Xi'an: China, 1997. 478～481

6 Tao Yang, et al. A Study of Throughput Probability of Statistical Multiplexer with Multi-port Buffer Scheme. 信息就是力量('97香港北京国际计算机会议论文集),清华大学出版社,1997. 153～156

7 Saluja K K, Anderson B D O. Easily diagnosable design at system level. Conf. on Comput. in Engin., Melbourne: Australia, 1981. 59～63

8 陶洋. 异种通信网互连控制关键技术研究:[重庆大学博士论文]. 1998. 6

$$P(d|C_k)=P(w_1,w_w,\cdots,w_n|C_k)=\prod_{j=1}^{n}P(w_j|C_k)$$

假定 N_k 表示训练样本集中属于第 C_k 类的邮件总数，N 表示训练样本集中的邮件总数，先验概率 $P(C_k)$ 为：

$$P(C_k)=N_k/N \qquad (2)$$

在计算条件概率中，为防止文档中单词出现次数少，导致分子或分母为 0 而使系统不能运行的情况，可以对朴素贝叶斯算法进行修正。假定 $N(w_i,d_j)$ 表示单词 w_i 出现在文档 d_j 的总数；$P(C_j|d_j)=\{0,1\}$，若训练集中的邮件 d_j 属于 C_j 类，则取 1，否则取 0；$|V|$ 表示向量空间模型中特征项的总数，修正后所得到的条件概率 $P(w_i|C_k)$ 公式为：

$$P(w_i|C_k)=\frac{1+\sum_{j=1}^{|V|}N(w_i,d_j)\times P(C_j|d_j)}{|V|+\sum_{i=1}^{|V|}\sum_{j=1}^{|D|}N(w_i,d_j)\times P(C_j|d_j)}$$

3. 基于最小风险朴素贝叶斯算法

由于在邮件的过滤过程中，朴素贝叶斯算法没有考虑合法邮件被错判为垃圾邮件的情况，因此，我们采用基于最小风险的贝叶斯算法进行邮件过滤。

在前述朴素贝叶斯算法的基础上，考虑在合法邮件被错判后所带来的风险或损失因素，设：

1)观察 ω 是 d 维随机向量 $\omega=[\omega_1,\omega_2,\cdots,\omega_d]^T$；

2)状态空间 Q 由 m 个自然状态(m 类)组成，$Q=\{C_1,C_2,\cdots,C_m\}$；

3)决策空间由 a 个决策 a_i，$i=1,2,\cdots,a$ 组成；

4)损失因子为 $\lambda(a_i,C_j)$，λ 表示当真实状态为 C_j 而所采取的决策为 a_i 时所带来的损失。

引入损失的概念后，在考虑错判所造成的损失时，不能只根据后验概率的大小来做决策，而必须考虑所采取的决策是否使损失最小。因此对于给定的一个邮件 d，如果采取决策 a_i，则其相应的条件期望损失为：

$$R(a_i|d)=\sum_{j=1}^{m}\lambda(a_i,C_j)\times P(C_j|d),\ i=1,2,\cdots,a$$

在考虑邮件错判时，希望损失最小，因此最小风险贝叶斯的判别规则为：如果 $R(a_i|d)=\min R(a_j|d)$，$j=1,2,\cdots,m$，则 d 属于 a_i 类。

在邮件过滤中，最小风险贝叶斯决策按如下所示步骤进行：

1)计算 $P(c_j)$，$P(d|c_j)$，$j=1,2$ 及待分类的邮件，根据贝叶斯公式计算出后验概率；

2)利用计算出来的后验概率及提供的决策表，按式(2)计算出采取 a_i，$i=1,2$ 的条件风险 $R(a_i|d)$；

3)对 2)中所得到的条件风险值 $R(a_i|d)$，$i=1,2$ 进行比较，找出使条件风险最小的决策 a_k，则 a_k 就是最小风险贝叶斯决策。

4. 实验结果

通过多种渠道我们收集了 400 篇邮件，其中垃圾邮件 310 篇，合法邮件 90 篇，这样的取样比例，倾向于把合法邮件误判为垃圾邮件。随后将邮件分为 4 组，随机选取其中三组邮件作为训练样本，对邮件进行测试分类(在实验中仅把邮件划分为正常邮件(C_1)与垃圾邮件(C_2)两类)。为避免实验的偶然性，重组训练样本集，对测试样本进行反复测试，取其平均值作为测试的结果。

利用向量空间模型计算出训练集中邮件的特征向量值，采用朴素贝叶斯算法，对邮件进行分类，得到表 1 所示结果。

表 1　基于朴素贝叶斯邮件过滤算法的实验结果

	系统评判为正常	系统评判为垃圾	总数
人工评判为正常	86	4	90
人工评判为垃圾	1.25	308.75	310
总数	87.25	312.75	400

从表 1 可知，对这 400 篇邮件进行测试分类，其中人工分类有 90 篇合法邮件和 310 篇垃圾邮件，而系统分类有 87.25 篇合法邮件和 312.75 篇垃圾邮件，有 4 篇合法邮件被系统误判为垃圾邮件，有 1.25 篇垃圾邮件被系统误判为合法邮件。

采用基于最小风险贝叶斯算法进行邮件过滤，表 2 提供每次测试过程中所需要的损失因子 λ，对邮件进行分类，得到表 3，通过调整 λ_{21} 阈值，对实验结果进行比较，选取合适的 λ_{21} 值，从而使合法邮件不被误判为垃圾邮件的同时，尽可能地提高分类的精确度。

表 2　决策表

次数	λ_{11}	λ_{12}	λ_{21}	λ_{22}
1	0	0.4	0.6	0
2	0	0.3	0.7	0
3	0	0.2	0.8	0
4	0	0.1	0.9	0

表 3　最小风险贝叶斯邮件过滤算法的实验结果

λ_{12}		系统评判为正常	系统评判为垃圾	总数
0.4	人工评判为正常	86.5	3.5	90
	人工评判为垃圾	1.25	308.75	310
	总数	87.75	312.25	400
0.3	人工评判为正常	87	3	90
	人工评判为垃圾	2.25	307.75	310
	总数	89.25	310.75	400
0.2	人工评判为正常	87	3	90
	人工评判为垃圾	2.5	307.5	310
	总数	89.5	310.5	400
0.1	人工评判为正常	87.5	2.5	90
	人工评判为垃圾	3.75	306.25	310
	总数	91.25	308.75	400

表 4　两种方法的查全率和查准率

算法 \ 比较内容		查全率	查准率
朴素贝叶斯算法		0.9555	0.9857
基于最小风险贝叶斯算法	$\lambda_{21}=0.6$	0.9611	0.9858
	$\lambda_{21}=0.7$	0.9667	0.9748
	$\lambda_{21}=0.8$	0.9667	0.9721
	$\lambda_{21}=0.9$	0.9722	0.9589

基于最小风险贝叶斯邮件过滤算法的实验结果如表 3 所示。从表 3 可知，当 λ_{21} 取 0.6 时，系统分类有 87.75 篇合法邮件和 312.25 篇垃圾邮件，有 3.5 篇合法邮件被系统误判为垃圾邮件，有 1.25 篇垃圾邮件被系统误判为合法邮件；当 λ_{21} 取 0.7 时，系统分类有 89.25 篇合法邮件和 310.75 篇垃圾邮件，有 3 篇合法邮件被系统误判为垃圾邮件，有 2.25 篇垃圾邮件被系统误判为合法邮件；当 λ_{21} 取 0.8 时，系统分类有

(下转第 46 页)

HTTP 请求和报头中有很多对负载平衡有用的信息，首先，我们可以从这些信息中获知客户端所请求的 URL 和网页，利用这个信息负载平衡设备就可以将所有的图像请求引导到一个图像服务器，或者根据 URL 的数据库查询内容调用 CGI 程序，将请求引导到一个专用的高性能数据库服务器。注意唯一能局限这些信息获取的因素是负载平衡设备本身的灵活程度。如果网络管理员熟悉 Web 内容交换技术，他可以仅仅根据 HTTP 报头的 cookie 字段来使用 Web 内容交换技术，改善对特定客户的服务，如果能从 HTTP 请求中找到一些规律，还可以充分利用它作出各种决策。除了 TCP 连接表的问题外，如何查找合适的 HTTP 报头信息以及作出负载平衡决策的过程，是影响 Web 内容交换技术性能的重要问题。

这种负载均衡技术依赖于特定的协议，因此其使用范围有限。

2.6 利用负载均衡设备的功能

目前，许多厂商推出了专用于平衡服务器负载的负载均衡器。目前负载均衡器生产商有：Intel、Alteon Web、Arrow Point、Coyote Point、F5 Networks、Foundry Networks、HydraWeb 以及 RADWare 等。所提供的解决方案主要包括如下几大方面：

1. 服务器群集的负载均衡，主要提供 Web Server Director，它可以监视所有的用户请求，并在可用的服务器群之间进行智能化的负荷分配，从而可以提供容错、冗余、优化和可扩展性能。

2. 高速缓存服务器的负载平衡，它可以提供优化的 Internet 访问和存储资源使用率，同时，也使整个服务器群的性能得以最大程度的发挥。

3. 防火墙的负载均衡，例如，RadWare 公司提供的 FireProof 是一种动态负载平衡系统，可有效地管理多个防火墙和其他安全设备上的流量。它使用的算法能够监视客户的数量和每个防火墙上的负载，并在各单元之间动态地平均分配流量，同时还可兼顾呼入和呼出的流量。

4. 链路的负载均衡，将多个线路的传输容量融合成一个单一的逻辑连接，适用于具有多个链接的网络。

在最新的负载均衡产品中，智能化越来越明显。一些智能化的负载均衡器能够侦测到像数据库错误、服务器不可用等信息，从而采取措施使会话恢复和重定向服务器，使电子商务能够得以顺利进行。另外多址负载均衡器可以对客户发来的访问请求进行解析，计算出最佳地址，然后将该地址返回客户，使客户自动连接到对其请求来说最佳的数据中心。

结束语 对于一个网络的负载均衡策略和技术的应用，从网络的不同层次可以采用不同的策略和技术，具体要依据网络的瓶颈所在。网络负载均衡技术的采用，使得自动故障恢复得以实现，服务的时间延长，24 小时×7 天可靠性和持续运行成为可能，同时负载均衡技术也为那些焦急等待大量数据和文件请求响应的客户提供了更快的响应时间。

参 考 文 献

1 沈中林，等. Linux 下集群虚拟服务器的研究与设计[j]. 计算机工程与应用，2000. 11

2 赵水宁，等. 多 Web 服务器负载均衡技术的研究[j]. 电信科学，2001. 7

3 姚耀文，等. 基于 Linux 的服务器群集方案[j]. 计算机工程，2001. 4

4 胡季敏，等. 使用动态负载均衡技术的 LINUX 高性能集群服务器研究[j]. 微型电脑应用，2001. 4

5 吕健，等. Linux 环境下的 Web 服务器负载均衡技术初探[j]. 科技情报开发与经济，2001. 2

（上接第 51 页）

89.5 篇合法邮件和 310.5 篇垃圾邮件，同样有 3 篇合法邮件被系统误判为垃圾邮件，有 2.5 篇垃圾邮件被系统误判为合法邮件；当 λ_{21} 取 0.9 时，系统分类有 91.25 篇合法邮件和 308.75 篇垃圾邮件，有 2.5 篇合法邮件被系统误判为垃圾邮件，有 3.75 篇垃圾邮件被系统误判为合法邮件。

从表 4 可知：当引入损失因子后，人工评判为正常而被系统评判为垃圾的邮件数都比贝叶斯算法所得到的结果小，说明合法邮件被误判为垃圾邮件的可能性减少，当 λ_{21} 取值增大时，查全率增加，但查准率下降。当 $\lambda_{21}=0.7$ 时，可得到较高查全率及查准率。

结束语 本文基于最小风险的贝叶斯方法提出一种新的邮件过滤算法，以减少合法邮件的误判率。实验结果表明，在邮件过滤中，与基于朴素贝叶斯的邮件过滤算法相比，基于最小风险的贝叶斯邮件过滤算法在查全率和查准率的性能方面有一定提高。当引入损失因子后查全率增加，合法邮件被系统误判为垃圾邮件的可能性减少，但同时，查准率有所下降，即垃圾邮件被系统误判为合法邮件的可能性增加，我们可利用合适的阈值来得到较高的查全率及查准率。

参 考 文 献

1 林亚平. 概率分析进化算法及其研究进展. 计算机研究与发展，2001，37(1)：43～49

2 Sahami M，et al. A Bayesian Approach to Filtering E-Mail. http://robotics. stanford. edu/users/sahami/papers-dir/. spam. ps，1998

3 Sahami M. Using Machine Learning to Improve Information Access：[PhD thesis]. Stanford University，Dec. 1998. 11～29，170～180

4 Rennie J D M. ifile：An Application of Machine Learning to E-Mail Filtering. http://www. cs. cmu. edu/～jr6b/papers/ifile98. ps，1998

5 Cohen W W. Learning rules that classify E-Mail. In：Proc. of the AAAI Spring Symposium on Machine Learning in Information Access，1996

6 Payne T. Learning Email Filtering Rules with Magi A Mail Agent Interface，MSc Thsis，University of Aberdeen，Scotland，1994

7 边肇祺，张学工等编著. 模式识别. 北京：清华大学出版社，第二版，1999. 9～16

8 Lewis D D. Feature Selection and Feature Extraction for Text Categorization. http://www. research. att. com/～lewis/chronobib. html/lewis92e. ps，1992

9 Nigam K，et al. Using EM to Classify Text from Labeled and Unlabeled Document. http://www-2. cs. emu. edu/pepole/mccallum/emcat-mlj2000. ps，1998

10 McCallum A，Nigam K. A Comparison of Event Model for Naive Bayes Text Classification. http://www-2. cs. cmu. edu/people/mccallum/multionmial-aaai98w. ps，1998