基于 Logistic 回归的中文垃圾邮件过滤方法*)

王庆幸[1] 徐从富[1] 何 俊[2]

(浙江大学计算机学院 杭州 310027)[1] (浙江省辐射环境监测站 杭州 310012)[2]

摘 要 研究如何实现 Logistic 回归模型在中文垃圾邮件过滤中的应用，给出了关键技术，并将其应用于 SEWM2007[1)] 垃圾邮件语料库上，取得了较优的过滤效果。还分析了影响正常邮件误判率、垃圾邮件误判率和精确率等因子。对比实验结果表明，应用于中文垃圾邮件过滤的 Logistic 回归模型与 SVM 相比具有更优的 ROC 指标和更快的运行效率。

关键词 垃圾邮件过滤，Logistic 回归，ham%，spam%，ROC

Filtering Chinese Spam Email Using Logistic Regression

WANG Qing-xing[1] XU Cong-fu[1] HE Jun[2]

(College of Computer Science, Zhejiang University, Hangzhou 310027, China)[1]

(Radiation Environment Monitoring Station of Zhejiang Province, Hangzhou 310012, China)[2]

Abstract We applied the logistic regression model to filter Chinese spam email, described the key approaches of our spam filter, and conduct experiments on SEWM2007 spam corpus. Additionally, we analyzed factors influencing the ham misclassification rate (ham%), the spam misclassification rate (spam%) and the accuracy of our filter. Compared with SVM, our filter is better in terms of receiver operating characteristics (ROC) and efficiency.

Keywords Spam filtering, Logistic regression, Ham%, Spam%, ROC

1 引言

随着垃圾邮件的泛滥，各种反垃圾邮件技术应运而生。当前，基于内容的智能邮件过滤技术已成为研究重点，支持向量机(Support Vector Machine, SVM)、动态马尔可夫建模(Dynamic Markov Modeling, DMM)、Winnow 等机器学习方法都已成功应用于邮件分类领域。这些方法的基本思路是，将垃圾邮件过滤看成一个两类问题，研究从样本邮件出发寻找规律(或分类器)，利用规律(或分类器)对未知邮件进行预测。

通常，可将机器学习技术划分为区分模型(Discriminative Model，如 Logistic 回归、SVM)和生成模型(Generative Model，如 Naïve Bayes)两类。Hulten 和 Goodman[1] 通过比较后认为，在 PU-1 垃圾邮件语料库上，区分模型的邮件分类效果要好于生成模型。本文作者在 SEWM2007 垃圾邮件过滤应用比赛中所得到的结果也充分证明了上述观点。特别值得一提的是，与其它代表队所采用的生成模型相比，本文作者所开发的基于 SVM 的垃圾邮件过滤器的性能是这次比赛中最为突出的[2)]。

然而，对于不同的区分模型，其性能也有差异。例如，Hsu, Chang 和 Lin 等人[2] 将 Logistic 回归应用于文本分类，在大数据集上取得了较线性 C-SVM 更优的精确率；Lynam 和 Cormack[3]，Goodman 和 Yih[4] 将 Logistic 回归应用于英文垃圾邮件分类，在 TREC 垃圾邮件语料库上进行了验证，效果较好。与 SVM 相比，我国学者研究 Logistic 回归在文本(特别是中文)分类等领域中的应用较少。本文主要研究如何实现 Logistic 回归模型在以中文为主的垃圾邮件(简称中文垃圾邮件)过滤中的应用，取得了较优的过滤效果。

2 Logistic 回归模型

回归分析是研究因变量 Y 与自变量 $X_1, X_2, \cdots, X_p$ 之间的关系，建立 Y 为 $X_1, X_2, \cdots, X_p$ 函数(回归函数)的过程，其主要目的是用自变量 $X_1, X_2, \cdots, X_p$ 对 Y 进行预测。Logistic 回归模型通过构建似然比(Likelihood ratio)的对数为一线性函数来实现，适合于连续与离散变量，具有较少的参数数量，可在宽范围的分布中应用[5]。

2.1 两类问题的 Logistic 回归模型

两类问题的 Logistic 回归模型是预测向量 $x=(x_1, x_2, \cdots, x_p)^T$ ($x_1, x_2, \cdots, x_p$ 为 $X_1, X_2, \cdots, X_p$ 的观测值)关于类 y ($y\in\{1,-1\}$)的后验概率。模型基于如下基本假设：类条件概率密度函数(Probability density function)的对数之差在 x 上是线性的，即

$$\log\left(\frac{p(x|y=1)}{p(x|y=-1)}\right)=\beta_0+\beta^T x \quad (1)$$

实践表明，上述模型在许多情况下都是正确的，并已大量应用于背离正态分布的真实数据集中[5]。根据贝叶斯定理，

*)本文研究受国家 863 计划项目(No. 2007AA01Z197)，国家自然科学基金项目(No. 60402010)的资助。**王庆幸** 工程硕士，主要研究方向为人工智能、信息处理等；**徐从富** 副教授，硕士生导师，主要研究方向为人工智能、数据挖掘、信息融合等；**何 俊** 高级工程师，主要研究方向为智能信息处理等。

1)第五届全国搜索引擎和网上信息挖掘学术研讨会(The 5th National Symposium of Search Engine and Web Mining(SEWM2007))。

2)实验结果可参见 http://www.cwirf.org/2007WebTrack/cct/SEWM07_SPAMoverview.ppt

式(2.1)的等价形式为

$$p(y=\pm 1|x)=\frac{1}{1+\exp(-y(\beta'_0+\beta^T x))} \tag{2}$$

其中，$\beta'_0=\beta_0+\log\frac{p(y=1)}{p(y=-1)}$。式(2)即为两类问题的 Logistic 回归函数，参数 β 称为权向量，可通过非线性优化方法进行估计。

2.2 参数估计

两类问题的 Logistic 回归模型的权向量可用包含向量 $\chi_1^{(1)},\cdots,\chi_m^{(1)}$（属于类 $y=1$）和向量 $\chi_1^{(-1)},\cdots,\chi_n^{(-1)}$（属于类 $y=-1$）的 $m+n$ 个训练集数据进行有监督学习，用最大似然法(Maximum likelihood)进行参数估计。下面给出基于公式(2)的对数似然方程(log-likelihood function)：

$$L=\sum_{k=1}^{m}\log(p(y=1|\chi_k^{(1)}))+\sum_{k=1}^{n}\log(p(y=-1|\chi_k^{(-1)})) \tag{3}$$

代入式(2)，有

$$L=\sum_{k=1}^{m}(\beta_0'+\beta^T\chi_k^{(1)})-\sum_{\text{所有}\chi}\log(1+\exp(-y(\beta_\beta+\beta^T\chi))) \tag{4}$$

求 L 最大，有下式成立：

$$\max_{\beta^T,\beta_0}L\equiv\min_{\beta^T,\beta_0}\sum_{\text{所有}\chi}\log(1+\exp(-y(\beta_0+\beta^T\chi))) \tag{5}$$

式(5)可用共轭梯度法、拟牛顿方法等无约束最优化方法求出使 L 达到局部最大的权向量。

为解决过拟合(Overfitting)问题，Lin[6]，Zhang[7]等通过正则化方法获得了更好的泛化能力。其中，Lin 等人给出 L2-regularized Logistic 模型[6]：

$$\min_w\{f(w)=\frac{1}{2}w^Tw+C\sum_{i=1}^{l}\log(1+\exp(-y_i(w^Tx_i)))\} \tag{6}$$

其中，C 为自定义的平衡参数。显然，在定义了扩展 $(p+1)$ 维的向量 $x\equiv[\chi^T,1]^T$ 和 $w\equiv[\beta^T,\beta_0]^T$，添加正则化项 $\frac{1}{2}w^Tw$ 后，式(6)可从式(5)得到。

3 关键技术

3.1 邮件解析

邮件文本报文的格式由 RFC822 定义，而声音、图像等 MIME 文档类型则由 RFC2045，2046 和 2047 定义。在基于内容的垃圾邮件过滤方法中，主要判定依据是邮件的主题(Subject)、文本主体(Data)以及附件(Attach)的名称，这些内容必须得到正确解析。

本文使用 JAVA Mail 的 API 对垃圾邮件公共语料库进行解析。将邮件实例化 MimeMessage 类的对象，通过 getSubject()方法提取邮件的主题；用 getContent()方法返回一个 MultiPartEncryped 对象，对该对象使用 tostring()方法得到 ASCII 或 ISO8859 格式的文本主体，使用 getContent()、getFileName()方法得到 MimeMultipart 等类型的文本主体和图像(Image)、应用程序(Application)的文件名。

3.2 分词

分词技术被广泛应用于搜索引擎、信息提取和机器翻译等领域，其实现方法可分为机械匹配法、基于规则的方法和基于统计的方法。

由于垃圾邮件经常采用文字变形，或加入干扰文字等方法来躲避过滤器的识别，这些特殊性在一定程度上弱化了因分词误差而导致的歧义性；垃圾邮件语言的多样性，要求所设计的分词方法能够兼容除中文以外的其它语言。

实验中，使用标点、空格等自然切分标志对邮件进行预处理，实现英文等非中文字符的切分，同时将邮件切分为更小的单位——句子，然后采用最大匹配法(Maximum Matching，MM)对预处理后的句子进行切分[3]。为提高分词效率，字典的加载使用了 HashMap 表。

3.3 特征提取与特征值计算

本文借助向量空间模型(Vector Space Model，VSM)[4]的思想，将邮件表示成加权的特征向量。

经分词后，邮件 e_j 被表示成一个集合 $\{t_1,\cdots,t_{|e|}\}$，$|e|$ 为邮件 e_j 中词的个数。集合中的每一个元素 t_k 作为一个特征，一个训练集 Ts（$Ts=\{e_1,\cdots,e_{|Ts|}\}$，$|Ts|$ 为训练集 Ts 中邮件的个数）的所有邮件的所有词组成一个特征空间 H（$H=\{t_1,\cdots,t_{|H|}\}$，$|H|$ 为特征空间中特征的个数），每封邮件都可表示成该特征空间的一个向量 $\chi_j=\{\chi_{1j},\cdots,\chi_{|H|j}\}$，显然，此向量的特点是高维且稀疏。

通常地，词的权重可通过 TF-IDF 公式计算[8]。在邮件分类应用中，本文增加了一个权重因子 w_m 来改变一些特殊词（如用户自定义的词、参与构建邮件主题的词、黄色或暴力等敏感词）的权重。

t_k 的特征值通过以下公式计算：

$$\chi_{kj}=\frac{\#(t_k,e_j)\times\log\frac{|Ts|}{\#Ts(t_k)}}{\sqrt{\sum_{s=1}^{|e|}(\#(t_s,e_j)\times\log\frac{|Ts|}{\#Ts(t_s)})^2}}\times w_m \tag{7}$$

其中，$\#(t_k,e_j)$ 为词 t_k 在邮件 e_j 中出现的次数；$\#Ts(t_k)$ 为训练集 Ts 出现词 t_k 的邮件个数。

3.4 信任域方法(Trust Region Method)

求解无约束最优化问题的信任域方法是通过一定的模式在信任域内优化目标函数的二次逼进式，直至满足精度要求的过程。给定一点 $x^{(k)}$，信任域通常取 $x^{(k)}$ 为中心的球域。考虑无约束问题[9]

$$\min f(x),x\in\mathcal{R} \tag{8}$$

将 $f(x)$ 在给定点 $x^{(k)}$ 展成 Taylor 级数，并取二次近似，得到

$$f(x)\approx f(x^{(k)})+\nabla f(x^{(k)})^{(T)}d+\frac{1}{2}d^T\nabla^2 f(x^{(k)})d$$

其中，$d=x-x^{(k)}$。给定可行点 $x^{(k)}$、信赖域半径 r_k，令 $\|d\|\leqslant r_k$，通过求解下列约束问题来求 $f(x)$ 极小点：

$$\min\{q_k(d)=f(x^{(k)})+\nabla f(x^{(k)})^{(T)}d+\frac{1}{2}d^T\nabla^2 f(x^{(k)})d\}$$

$$s.t.\ \|d\|\leqslant r_k \tag{9}$$

Lin[6]等人将式(9)表示为

$$\min\{q_k(d)=\nabla f(x^{(k)})^Td+\frac{1}{2}d^T\nabla^2 f(x^{(k)})d\}$$

$$s.t.\ \|d\|\leqslant r_k \tag{10}$$

通过考虑约束条件 $\|d\|\leqslant r_k$ 的共轭梯度法求出式(10)的最优解 d^k，然后根据函数值实际下降量与预测量之比来确定 $x^{(k+1)}$ 和 r_{k+1} 的迭代规则，由函数的梯度值判定 $x^{(k+1)}$ 能否

3)使用了北大天王的中文字典。http://net.pku.edu.cn/~webg/src/ChSeg/words.dict.

4)Salton 等人于 20 世纪 60 年代末提出的一种信息检索方法。

作为式(8)的近似解。

3.5 训练与测试

一组邮件被表示成特征空间的一组向量以后，放入两类问题的 Logistic 回归模型进行训练，估算出回归模型的权向量，就可以对用向量表示的邮件进行分类预测，测试的过程就是求出该邮件属于某一类的后验概率。训练和测试使用 LIBLINEAR(Lin, Weng, and Keerthi 2007)的 L2-regularized Logistic 模型代码[5]。

4 实验

4.1 数据集

SEWM2007 公共垃圾邮件语料库提供了 60000 份邮件，主要为中文邮件，其中垃圾邮件 45000 封，正常邮件 15000 封。垃圾邮件来源于校园网垃圾邮件过滤系统过滤下来的垃圾邮件和用户报告的垃圾邮件，从 234592 个垃圾邮件样本中选出；正常邮件部分为志愿者提供，部分通过使用真实邮件的头信息和 Web 上抓取邮件内容进行合成，并考虑了与垃圾邮件特征词的适当交叉。为保护志愿者隐私，语料中去除了 IP 地址、服务器名、用户名等信息[10]。

4.2 训练集的划分

本文以 5:5 的比例将数据集随机划分为 2 个子集 A 和 B，如表 1 所示[6]。首先，在子集 A 上训练，在子集 B 上测试；然后交换子集进行训练和测试，即在 B 上训练，在 A 上测试；用两次训练、测试的平均值作为终值。这种轮转法与保持法相比，可以减少偏差，与交叉验证法相比计算量较少，是保持法和交叉验证法的折衷。

表 1 训练集的划分

|Ts|	Set	spam	ham	spam:ham	|H|
58525	A	21777	7488	2.908	85816
	B	21750	7510	2.896	90082

4.3 评价体系

垃圾邮件过滤的性能评价可以借用文本分类的评价指标。本文采用 ham%，spam%和 Accuracy 等指标来评价垃圾邮件过滤的判别能力。考虑到在实际应用中，人们宁愿接收多一点的垃圾邮件，也不愿正常邮件被误判为垃圾邮件，因此，我们将 ham%值作为评价垃圾邮件过滤性能的第一依据，spam%，Accuracy 等作为参考。

4.4 结果及其分析

表 2 不同的平衡参数 C 对性能的影响($w_m=1:3, w_i=1:3$)

C	ham%(%)	spam%(%)	Accuracy(%)
0.25	4.070(3.93-4.21)	0.835(0.86-0.81)	98.335(98.35-98.32)
1	2.920(2.78-3.06)	0.570(0.59-0.55)	98.830(98.85-98.81)
4	2.435(2.21-2.66)	0.475(0.5-0.45)	99.025(99.06-98.99)
25	2.125(2.18-2.32)	0.430(0.46-0.4)	99.095(99.09-99.10)
64	2.120(1.98-2.27)	0.435(0.48-0.39)	99.135(99.14-99.13)
125	2.075(1.97-2.18)	0.445(0.49-0.4)	99.140(99.13-99.15)
256	2.070(1.98-2.16)	0.440(0.49-0.39)	99.140(99.13-99.15)
625	2.120(1.93-2.31)	0.450(0.51-0.39)	99.075(99.03-99.12)

本文选择了 3 组不同的词权重因子 w_m 用于特征值计算，6 种不同的平衡参数 C 和 3 种类间权重因子 w_i 参与训练，以观察它们对垃圾邮件过滤的性能影响。

由表 2 可知，平衡参数 C 能明显改善垃圾邮件过滤的性能，在 $C=256$ 时，ham%与 Accuracy 同时达到最优。

在表 3 和表 4 中，随着词权重因子 w_m、类间权重因子 w_i 的增加，ham%指标改善的同时 spam%有所下降。

表 3 词权重因子 w_m 对性能的影响($C=256, w_i=1:3$)

w_m	ham%(%)	spam%(%)	Accuracy(%)
1:1	2.770(2.72-2.82)	0.250(0.23-0.27)	98.925(99.13-99.08)
1:3	2.070(1.98-2.16)	0.440(0.49-0.39)	99.140(99.13-99.15)
1:5	2.280(2.01-2.55)	0.660(0.73-0.59)	99.105(98.94-98.91)

表 4 类间权重因子 w_i 对性能的影响($C=256, w_m=1:3$)

w_i	ham%(%)	spam%(%)	Accuracy(%)
1:1	2.920(2.62-3.22)	0.315(0.33-0.3)	99.015(99.08-98.95)
1:3	2.070(1.98-2.16)	0.440(0.49-0.39)	99.140(99.13-99.15)
1:5	1.890(1.84-1.94)	0.515(0.59-0.44)	99.135(99.09-99.18)

表 5 反映的是词权重因子 w_m 和类间权重因子 w_i 的联合影响，可以看出类间权重因子 w_i 对垃圾邮件过滤性能的影响程度大于词权重因子 w_m。

表 5 词权重因子 w_m、类间权重因子 w_i 对性能的联合影响($C=256$)

w_i	w_m	ham%(%)	spam%(%)	Accuracy(%)
1:5	1:3	1.890(1.84—1.94)	0.515(0.59-0.44)	99.135(99.09-99.18)
1:5	1:5	1.940(1.77-2.11)	0.650(0.70-0.60)	99.015(99.02-99.01)
1:3	1:3	2.070(1.98-2.16)	0.440(0.49-0.39)	99.140(99.13-99.15)
1:3	1:5	2.280(2.01-2.55)	0.66(0.73-0.59)	98.925(98.94-98.91)
1:5	1:1	2.420(2.38-2.46)	0.315(0.33-0.3)	99.150(99.15-99.15)

为了进一步衡量 Logistic 回归模型过滤垃圾邮件的性能，本文将 Logistic 回归模型与 SVM 进行了对比[7]。我们选择了分别使 Logistic 回归模型和 SVM 达到邮件过滤性能最优的 C，w_i 和 w_m 因子，以 ham%为横坐标，以 spam%为纵坐标，取不同的阈值，画出 ROC (Receiver operating characteristics)曲线(如图 1 所示)。ROC 曲线下方的面积(Area under the ROC curve, AUC)反映了邮件过滤效率的一个累计度量。显然，Logistic 回归模型的表现要胜于 SVM。

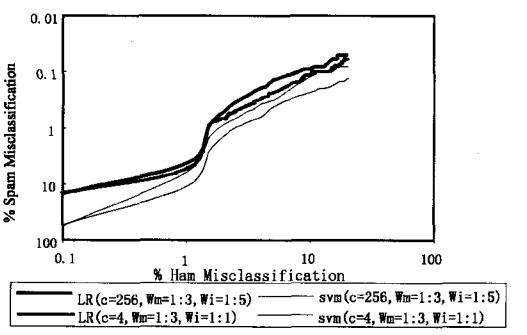

图 1 Logistic 回归与 SVM 的 ROC 比较

表 6 给出了上述两种模型的训练和预测时间。对比结果表明，对于垃圾邮件过滤这种样本数目和特征数目都很巨大

(下转第 229 页)

5) Software available at http://www.csie.ntu.edu.tw/~cjlin/liblr

6)SEWM200 公共垃圾邮件语料库经邮件解析后有约 58525 封邮件得到结果，参与了本次训练和测试实验。经统计，剩余的 1475 封邮件中正常邮件为 2 份。

7)使用 Libsvm. Software availbale at http://www.csie.ntu.edu.te/~cjlin/libsvm

[11] Vasic B, Pedagani K. Run-length-limited low-density parity check codes based on deliberate error insertion. IEEE Trans. Magnetics, 2004, 40(3): 1738-1743

[12] Peel C M, Pegram S, Mcmahon A T. Global analysis of runs of annual precipitation and runoff equal to or below the median: run length. International Journal of Climatology, 2004, 24(7): 807-822

[13] Nagy Z, Zeger K. Bit-stuffing algorithms and analysis for run-length constrained channels in two and three dimensions. IEEE Trans. Information Theory, 2004, 50(12): 3146-3169

[14] Makinen V, Navarro G, Ukkonen E. Approximate matching of run-length compressed strings. Algorithmica, 2003, 35(4): 347-369

[15] Radson D, Boyd H A. Graphical representation of run length distributions. Quality Engineering, 2005, 17(2): 301-308

[16] Amir A, Landau MG, Sokol D. Inplace run-length 2 d compressed search. Theoretical Computer Science, 2003, 290(3): 1361-1383

[17] Monasse P, Guichard F. Fast computation of a contrast-invariant image representation. IEEE Trans. Image Processing, 2000, 9(5): 860-872

[18] Flusser J. Refined moment calculation using image block representation. IEEE Trans. Image Processing, 2000, 9(11): 1977-1978

[19] Voronin V. Holographic representation in image processing tasks. Pattern Recognition and Image Analysis, 2001, 11(1): 265-267

[20] Liu Y, Ranganath S, Zhou X. Wavelet-based image segment representation. Electronics Letters, 2002, 38(19): 1091-1092

[21] Kharinov V M. Representation of image information for machine computation. Pattern Recognition and Image Analysis, 2005, 15(1): 212-214

[22] Malo J, Epifanio I, Rafael N, et al. Nonlinear image representation for efficient perceptual coding. IEEE Trans. Image Process, 2006, 15 (1): 68-80

[23] Samet H. Data structures for quadtree approximation and compression. Communications of the ACM, 1985, 28(9): 973-993

[24] Samet H, Webber R E. Storing a collection of polygons using quadtrees. ACM Trans. on Graphics, 1985, 4(3): 182-222

[25] Li S X, Loew M H. The quadcode and its arithmetic. Communications of the ACM, 1987, 30(7): 621-626

[26] Gargantini I. An effective way to represent quadtrees. Communications of the ACM, 1982, 25(12): 905-910

[27] Zheng Y P, Chen C B. Study on a NAM-based color image representation method. Journal of Software, 2007, 18(11): 2932-2941

[28] Chen C B, Hu W J, Wan L. Direct non-symmetry and anti-packing pattern representation model of medical images // Proceedings of the First International Conference on Bioinformatics and Biomedical Engineering. Wuhan, China, July2007: 1011-1018

[29] Zheng Y P, Chen C B, Sarem M. A novel algorithm for triangle non-symmetry and anti-packing pattern representation model of gray images // Proceedings of the Third International Conference on Intelligent Computing ICIC'07, LNCS 4681. Qingdao, China, August 2007: 832-841

(上接第 199 页)

的应用场合下，实验中使用的 Logistic 回归模型及其参数估计方法较 SVM 有更快的算法收敛速度、更高的运行效率。

表 6 Logistic 回归与 SVM 运行效率的比较

Model	Train Time(s)	Test Time(s)	Total(s)
LR(C=4, w_i=1:3, w_m=1:1)	34(30-38)	16(13-18)	50(43-56)
LR(C=256, w_i=1:3, w_m=1:5)	43(39-46)	16(17-15)	59(56-61)
SVM(C=4, w_i=1:3, w_m=1:1)	2115 (1858-2372)	157 (127-186)	2272 (1985-2558)
SVM(C=256, w_i=1:3, w_m=1:5)	1548 (1460-1635)	115 (111-118)	1663 (1571-1753)

结束语 本文实验结果表明，Logistic 回归模型具有较少的调节参数，能够在中文垃圾邮件过滤应用中取得很好的分类效果。与 SVM 相比，无论是在 ROC 分类指标上，还是在运行效率上，Logistic 回归模型都要优于后者。未来的主要工作是，如何将 Logistic 回归模型应用于在线中文垃圾邮件过滤系统中，并验证其效果。

参 考 文 献

[1] Hulten G, Goodman J. Tutorial on junk e-mail filtering // Proceedings of 21st International Conference on Machine Learning (ICML'2004). Banff, Canada, July 2004: 45-57. http://www.research.microsoft.com/~joshuago/tutorialOnJunkMailFilteringjune4.pdf

[2] Hsu C, Chang C, Lin C. A practical guide to support vector classification, 2007. http://www.csie.ntu.edu.tw/~cjlin

[3] Lynam T R, Cormack G V. On-line spam filter fusion // Proceedings of 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR'2006). Seattle, Washington, August 2006

[4] Goodman J, Yih W. Online discriminative spam Filter training // Proceedings of 3rd Conference on Email and Anti-Spam (CEAS'2006). July 2006: 113—116. http://www.ceas.cc/2006/allpapers.pdf

[5] [英]Andrew R W. 统计模式识别(第二版). 王萍，等译. 北京：电子工业出版社，2004

[6] Lin C, Weng R C, Keerthi S S. Trust region Newton method for large-scale logistic regression // Proceedings of 24th International Conference on Machine Learning (ICML'2007). June 2007. http://www.machinelearning.org/ proceedings/icml2007/papers/114.pdf

[7] Zhang J, Yang Y. Robustness of regularized linear classification methods in text categorization // Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR'2003). Toronto, Canada, July 28-August 1, 2003: 191-192. http://net.pku.edu.cn/~wbia/2004/ public_html/Readings/tmp/reading-2/Robustness%20of%20Regularized%20Linear%20Classification%20Methods%20in%20Text%20Categorization.pdf

[8] 陈宝林. 最优化理论与算法(第 2 版). 北京：清华大学出版社，2005

[9] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1-47

[10] 华南理工大学信息网络工程研究中心和广东省计算机网络重点实验室邮件评测小组. SEWM2007 垃圾邮件过滤系统评测，2007, 3. http://www.cwirf.org/ 2007WebTrack/cct/SEWM07_SPAMoverview.ppt