浙江科技学院学报,第 22 卷第 6 期,2010 年 12 月
Journal of Zhejiang University of Science and Technology
Vol. 22 No. 6, Dec. 2010

DOI: 10.3969/j.issn.1671-8798.2010.06.008

基于 HTML 卡方算法的垃圾邮件过滤器设计

孔　颖
(浙江科技学院 信息与电子工程学院,杭州 310023)

摘　要: 介绍基于 HTML 标签的卡方分布算法在垃圾邮件过滤中的应用。首先对通过浏览器收集到的邮件进行分析,将其转换为 HTML 源代码的形式,再根据 HTML 语言的特点对其进行特征提取,从而达到邮件预处理的目的。随后采用 LVQ 神经网络建立分类器模型,以达到最终分离正常邮件(ham)和垃圾邮件(spam)的目的。对比实验表明,结合 HTML 代码的卡方分布特征提取和 LVQ 神经网络的分类器模型效果更好。

关键词: 垃圾邮件过滤;HTML 标签;卡方分布
中图分类号: TP393.098　　**文献标识码**: A　　**文章编号**: 1671-8798(2010)06-0525-05

Design of spam filtering model based on HTML chi-square algorithm

KONG Ying
(School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China)

Abstract: We introduce the application of chi-square distribution in spam filtering based on HTML tag algorithm. Firstly, we analyze the contents of the e-mail and convert them into forms of HTML source code. Then we do feature extraction according to HTML language characteristics so as to achieve the purpose of e-mail pre-processing. Finally, we use LVQ neural networks to design a classifier model in order to realize the purpose of separating ham from spam. The comparison test results show that the LVQ neural network classifier based on HTML chi-square distribution has achieved better results.

Key words: spam filtering; HTML source tag; chi-square

电子邮件作为一种高效、经济的现代通信技术手段,已成为互联网最大的应用项目之一。然而,随之产生的垃圾邮件像瘟疫一样蔓延,污染网络环境,占用大量传输、存储和运算资源,影响了网络的正常运行,严重干扰了人们的正常生活,浪费用户的时间、精力,甚至造成很多额外的经济支出和信息安全隐患。垃圾邮件的判定和邮件的接收者有很大关系,不同用户对同一邮件的判断结果可能会存在差异。在《中国

收稿日期: 2010-04-12
作者简介: 孔　颖(1980—　),女,浙江杭州人,讲师,硕士,主要从事人工智能、机器学习、数据挖掘研究。

互联网协会反垃圾邮件规范》中，将垃圾邮件定义为具备如下部分或全部特征的电子邮件：

1）收件人事先没有提出要求或同意接收的广告、电子刊物、各种形式的宣传品等宣传性的电子邮件；

2）收件人无法拒收的电子邮件；

3）隐藏发件人身份、地址、标题等信息的电子邮件；

4）含有虚假的信息源、发件人、路由等信息的电子邮件；

5）含有病毒、恶意代码、色情、反动等不良信息或有害信息的电子邮件。

基于内容的机器学习判别方法是当前解决垃圾邮件问题的主流技术之一，包括 Ripper、决策树方法、Rough Set 方法等基于规则的方法，和 Bayes、SVM、动态马尔可夫建模(Dynamic Markov Modeling，DMM)、Winnow 等基于概率统计的方法[1-2]。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。随着人工智能、计算机技术的创新和发展，这种将机器学习方法应用于邮件分类领域一直成为当前研究的热点和重点。

本文在简要介绍基于 HTML 的卡方特征提取算法的基础上，将卡方特征提取和 HTML 标签相结合，通过神经网络建立模型，对邮件进行预测，得到了较好的实验结果。

1 邮件特征提取

1.1 HTML 标签分类

首先需要将收集到的邮件转换为 HTML 形式，这样可以将邮件特征使用标签来表示。由图 1 可以看到一封浏览器接收到的邮件示例。

将图 1 所示的邮件转换为 HTML 格式，由于文本文件的通用性，所有的邮件均可转换为 HTML 代码。而 HTML 为各类标签所组成，对于不同的标签而言，其所代表的含义各不相同，当然，处于标签中的内容所代表的含义也不同，表 1 大致给出了基本标签所代表的含义。

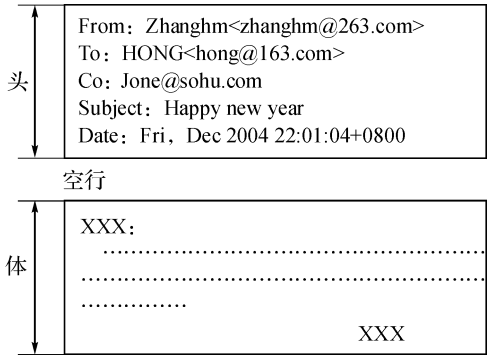

图 1 邮件示例图

Fig. 1 An example of an e-mail

表 1 基本标签含义表

Table 1 Meanings of basic tags

标签	含义
＜! doctype＞	Defines the document type
＜html＞	Defines an html document
＜body＞	Defines the body element
＜h1＞to＜h6＞	Defines header 1 to header 6
＜p＞	Defines a paragraph
＜br＞	Inserts a single line break
＜hr＞	Defines a horizontal rule
＜! ——…——＞	Defines a comment

当然，标签中许多都是用于格式控制的，对于文本内容没有任何影响，不会使文本特征丢失。将邮件转换格式后为了便于进一步进行特征提取，需要对某些细节进行一些处理，其中主要的方式为：去除含有图片的标签；去除超链接的标签；不考虑标签中与附件有关的内容；将特殊符号转化为特殊的标签。

完成了细节处理后，需要对各种标签进行分类，以便于对不同标签内的内容赋予不同的权值，达到最终的特征提取目的。根据 HTML 语言的特点，在标签分类时使用如表 2 所示的方式。

特征提取将标签中最为相关的特征集中在一个数据集中，考虑到整个数据集包含所有邮件的特征，如邮件中的单词、图片和它们在预处理时所产生的标签等[3]。由于特征提取广泛用于文本的分类，通常这些分类方法也可以用于处理垃圾邮件，在本实验中采用了 2 种提取法：TF-IDF 算法和卡方分布算法。

1.2 卡方特征提取

卡方分布处理的是某个特征的度与整个数据集之间的关系[4]。如 w 是数据集 C(由两部分构成)中的

一个特征,则 w 的卡方值可用式(1) 给出

$$x^2(w) = P(\text{spam}) \cdot x^2(w,\text{spam}) + P(\text{ham}) \cdot x^2(w,\text{ham}) \tag{1}$$

式(1) 中 $P(\text{spam})$ 和 $P(\text{ham})$ 分别表示垃圾邮件和正常邮件在数据集中出现的概率,这样就可以给出特征 w 在整个数据集中的卡方分布,如式(2) 所示:

$$x^2(w,c) = \frac{N(kn-ml)^2}{(k+m)(l+n)(k+l)(m+n)} \tag{2}$$

式(2) 中:k 为正常邮件数据集中包含特征 w 的邮件数量;l 为垃圾邮件中包含特征 w 的邮件数量;m 为正常邮件数据集中不包含特征 w 的邮件数量;n 为垃圾邮件数据集中不包含特征 w 的邮件数量;N 为正常邮件数据集中所有邮件的数量。同样的,所有的特征的卡方值均取其最高值,最终这些值均作为神经网络中的一个节点。

表 2　标签相关特征

Table 2　Characteristics of tags

标签	类别	标签	类别
a	3	html	2
abbr	2	i	2
acronym	2	img	3
b	2	input	3
base	3	ins	2
blockquote	3	kbd	2
body	2	label	2
br	2	li	2
button	3	map	3
caption	2	marquee	1
col	2	ol	2
comment	1	optin	2
del	2	p	2
em	2	select	2
font	3	style	1

1.3　TF-IDF 加权算法

目前加权使用最广泛的算法 ——TF-IDF 加权算法[5-6]:

$$W(t,\vec{d}) = tf(t,\vec{d}) \times \log\left(\frac{N}{n_t}\right) \tag{3}$$

式(3) 中,$W(t,d)$ 为特征项 t 在邮件 d 中的权重;$tf(t,d)$ 为特征项 t 在邮件内容中的词频;N 为训练文本的总数;n_t 为训练邮件集中出现特征项 t 的邮件数。用 TF-IDF 算法来计算特征词的权重值是表示当一个词在这篇邮件中出现的频率越高,同时在其他文档中出现的次数越少,则表明该词对于表示这篇文档的区分能力越强,所以其权重值就应该越大[7]。将所有词的权值排序,根据需要选择特征项。

为消除文档长度不一对文本表示方式的可能影响,往往需要对加权后的向量进行规范化处理,使得权值落在[0,1] 中。即:

$$W(t,\vec{d})^t = \frac{W(t,\vec{d})}{\sqrt{\sum_{i=1}^{n} W_i^2}} \tag{4}$$

1.4　基于 LVQ 神经网络的邮件分类器设计

LVQ 神经网络是一类混合神经网络,它分为有人值守和无人值守[7]。本实验中将模型分为两层,第一层是竞争层,该层中每个节点表示一个子集;而第二层为输出层,每一个节点均为一个集。每个集可以划分为若干个子集。由于 LVQ 神经网络可以通过结合不同的子集创造复杂的界限,故适合于将垃圾邮件从若干不同的子集中分辨出来。以下为筛选算法:

1) 初始化向量权重 $\boldsymbol{W} = \{W_1, W_2, \cdots, W_n\}$,学习率 $\alpha \in [0,1]$。

2) 从训练邮件集中选取一个示例,计算它各个向量之间的距离,分别取欧几里德距离和余弦距离,这些可以表示不同文本之间的相似性[8]。其中余弦距离可用式(5) 表示:

$$\text{sim}(u,v) = \frac{\sum_{k=1}^{n} w_{uk} w_{vk}}{\sqrt{\sum_{k=1}^{n} w_{uk}^2}} \cdot \frac{1}{\sqrt{\sum_{k=1}^{n} w_{vk}^2}} \tag{5}$$

3) 比较不同权向量之间的距离,在结果中,神经元之间最为相似的取值 1,其余隐藏层的输出层取值 0。

$$a^1 = \max(\mathrm{sim}(x, x^i)) \tag{6}$$

4）调整距离，如果一个输入示例属于数据集 r，那么在数据集 s 中神经元 c 拥有最大的权值，然后根据式(7)调整各个权值：

$$\begin{cases} w_c(t+1) = w_c(t) + u(t)[x(t) - w_c(t)]; r = s \\ w_c(t+1) = w_c(t) - u(t)[x(t) - w_c(t)]; r \neq s \\ w_i(t+1) = w_i(t); i \neq c \end{cases} \tag{7}$$

5）修改学习率 $U(t)$，当重复增加时降低 $U(t)$。

6）检查停止状况，并确认重复的次数足够多。

使用上述方法提取特征之后代入 LVQ 神经网络进行计算，使用 MATLAB可以模拟出如图 2 的 LVQ 神经网络。

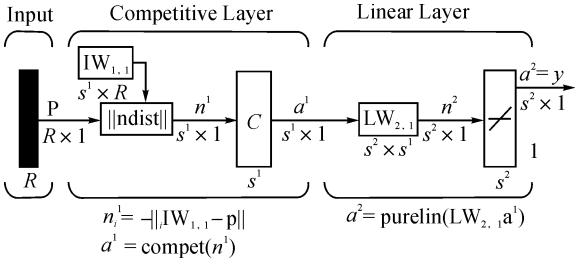

图 2 LVQ 神经网络

Fig. 2 LVQ neural network

1.5 过滤原理

邮件预处理是建立在训练模型的基础上的，因为用 LVQ 神经网络建立的模型是建立在许多学习样本邮件基础上的，需要巨大的计算资源。因此，用于建模的邮件如何进行预处理十分关键。

在特征提取过程中，通过把普通的邮件转化成 HTML 标签形式，减少分类过程中产生的特征向量数，把每个特征所出现的概率用到卡方算法中，最后再通过 LVQ 神经网络建立模型，进行邮件分类。电子邮件过滤模型如图 3 所示。

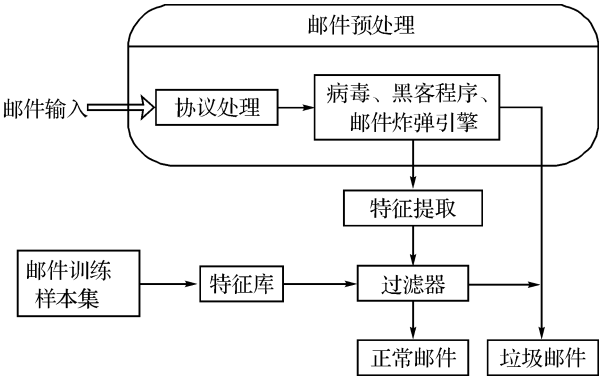

图 3 电子邮件过滤模型

Fig. 3 Filtering model of an e-mail

2 实验结果分析

评价一个解决分类问题的模型是否适用的一个直接手段就是看它的错分率，即错误分类数与总记录数的比值。所示必须测量用来建立模型和在建立模型过程中没有用到的记录组成的测试样本，选择最具有普遍意义的而不是最适合训练样本的模型。

考虑 N 封待测试邮件（N_s 封垃圾邮件和 N_h 封垃圾邮件，$N = N_s + N_h$），在算法邮件分类模型中，垃圾邮件被分类器正确判定的有 A 封，误判的有 B 封；正常邮件被分类器正确判定的有 C 封，误判的有 D 封，显然 $N_s = A + B$，$N_h = C + D$。根据定义，有下列各式成立：

$$\mathrm{ham} = \frac{C}{(C+D)} \times 100\%$$

$$\mathrm{spam} = \frac{A}{(A+B)} \times 100\%$$

$$\mathrm{Accuracy} = \frac{(A+C)}{N} \times 100\% \tag{8}$$

垃圾邮件过滤应用模型采用 ham%、spam%和 Accuracy 等传统分类指标，来分析特征提取和特征值计算方法、训练模型的选择对邮件分类模型的影响。

本次实验使用的是 SEWM2008 比赛中的数据集作为评测数据集邮件样本。抽取样本中的 4 000 封，其中正常邮件有 3 120 封，垃圾邮件有 880 封。分别用基于 HTML 的卡方和 TF-IDF 两种不同的特征提取方法，把得到的邮件特征向量通过 LVQ 神经网络模型进行过滤，从而得到的实验结果，如表 3 所示。

表 3　各实验结果比较

Table 3　Comparison of experimental results

邮件分类	邮件总数	卡方特征提取					TF-IDF 特征提取				
		判为(0)	判为(1)	正确率%	正确率平均%	准确率%	判为(0)	判为(1)	正确率%	正确率平均%	准确率%
垃圾邮件(0)	880	848	32	96.36	96.10	95.95	840	40	95.45	95.15	94.98
正常邮件(1)	3 120	130	2 990	95.83			161	2 959	94.84		
合计	4 000	978	3 022				1 001	2 999			

从表 3 实验结果的比较可以看出,在数据集足够大时,采用 LVQ 神经网络的分类器对于不同方法提取的数据集均有较好的结果,其中采用卡方分布法提取的数据集在处理结果方面略微优于传统的 TF-IDF 提取的数据集。实验还表明,不管是正常邮件分类、垃圾邮件分类还是整体分类,都具有较高的准确率。

3　结　语

从基于 HTML 的卡方特征提取方法和 LVQ 神经网络分类器结果可以看出,该模型是一种较好的垃圾邮件处理系统。通过转换邮件文本为 HTML 代码,便于处理其中内容,而使用 LVQ 神经网络的分类器,在数据集足够大时所得结果往往优于同等情况下的其他分类器,这在实际应用时具有一定的参考价值。

参考文献:

[1] YIH W T, MCCANN R, KOLCZ A. Improving spam filtering by detecting gray mail[C]//Fourth Conference on Email and Anti-Sparn. Mountain View, CA: CEAS,2007.

[2] CLEARY J G, WRITTEN I H. Data compressing using adaptive coding and partial string matching [J]. IEEE Transaction on Communications,1984,32(4):396-402.

[3] ZEITOUN I K, YEH L. Join indices as a tool for spatial data mining[C]//International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, Lecture Notes in Artificial Intelligence. Paris: Springer Press,2007:102-114.

[4] 刘洋,杜孝平,罗平,等. 垃圾邮件的智能分析、过滤及 Rough 集讨论[R]. 武汉:第十二届中国计算机学会网络与数据通信学术会议,2002.

[5] 王斌,潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报,2005,19(5):1-10.

[6] 程红蓉,秦志光,万明成,等. 图像垃圾邮件中文本区域的自动提取方法[J]. 解放军理工大学学报:自然科学版,2009,10(3):258-261.

[7] 王龙,李晓光,钟绍春. 基于 K-近邻法及移动 AGENT 技术的垃圾邮件检测系统研究[J]. 计算机应用研究,2009,26(7):2630-2632.

[8] 万明成,耿技,程红蓉,等. 图像型垃圾邮件过滤技术综述[J]. 计算机应用研究,2008,25(9):2579-2582.