

# 基于 Winnow 算法的反垃圾邮件引擎的设计与实现

张丽, 黄东

(东南大学自动控制系, 江苏南京 210096)

**摘要:**电子邮件(e-mail)是人们日常生活中不可缺少的通信手段之一,但是垃圾邮件却给人们带来了很大的危害。文中主要是针对中文垃圾邮件,给出了一种基于 Winnow 算法的基于邮件内容的反垃圾邮件引擎原型的设计,对于未知邮件可以达到较好的区分效果。首先对邮件的内容进行解码、分词,采用信息增益选取特征项;然后采用 Winnow 算法构造分类器;最后采用部分邮件样本进行测试,测试结果可以进行反馈学习。最后的测试数据分析表明系统达到了比较好的效果。

**关键词:**垃圾邮件;文本分类;特征选择;特征提取;Winnow 算法;反馈

**中图分类号:**TP393.098

**文献标识码:**A

**文章编号:**1005-3751(2006)04-0170-03

## Design and Implementation of One Prototype of Anti-Spam Engine Based on Winnow Algorithm

ZHANG Li, HUANG Dong

(Automation and Control Department of Southeast University, Nanjing 210096, China)

**Abstract:** Email is one of indispensable communication ways in daily life, but spam has done serious harm to people. In this paper present the design of an anti-spam engine based on Winnow algorithm and focus on Chinese spam, and the result of distinguishing from unknown mail is good. Firstly it decodes content of the mail, segments, and chooses terms with information gain. Then it constructs the classification. Finally it tests the result with partly mails, and the wrong will result in feedback study. The test data analysis shows that the system outcome is good.

**Key words:** spam; text categorization; feature selection; feature extraction; Winnow algorithm; feedback

### 0 引言

随着国际互联网 Internet 的发展和普及,电子邮件(e-mail)以其方便、快捷、低成本的独特魅力成为人们日常生活中不可缺少的通信手段之一。但电子邮件给人们带来极大便利的同时,也日益显示出其负面影响,那就是垃圾邮件。垃圾邮件已经给人们的日常生活带来了很大的危害。从电子邮件的结构出发,寻找垃圾邮件的特征,在发件人、收件人、邮件头、邮件正文内容等各方面展开邮件过滤工作,是垃圾邮件过滤常采用的基本方法。基于内容的垃圾邮件过滤技术最终归结是文本分类的问题,文中主要针对中文垃圾邮件,给出了一种基于 Winnow 算法的反垃圾邮件引擎设计和实现。

### 1 系统设计

文本分类是指在给定的分类体系下,根据文本的内容确定文本相关类别的过程。从数学角度看,文本分类是一个映射过程,即未标明类别的输入文本到给定分类的

一对一或者一对多的映射。文本分类技术已经在搜索引擎、邮件分类、信息过滤、防火墙等领域得到了广泛的应用。

从文本分类的角度看,垃圾邮件的过滤就是未知邮件的分类,归为垃圾邮件或者非垃圾邮件,属于二值分类问题。图1给出了系统的结构流程图。

本系统主要由3个模块组成,分别对应样本邮件的训练即分类器的构造和新邮件的分类以及分类结果的反馈。其中训练过程具体分为邮件预处理、特征选择和提取、邮件特征表示、选择分类算法训练最终生成分类器;分类过程分为邮件预处理、邮件特征表示,然后应用已有的分类器选取分类算法对新邮件分类。

#### 1.1 邮件的预处理

邮件的预处理包括邮件解码和自动分词。解码是按照相关 RFC 协议来进行一个编码的逆过程——解码,以得到邮件的实际内容。

汉语的书写以汉字作为基本单位,词与词之间没有明显的形态界限,要进行中文邮件的处理,必须首先将汉语的词与词分割开,即分词。分词的目的就是将邮件中的内容分为独立的可以表示信息的词<sup>[1]</sup>。自动分词是垃圾邮件处理的基础步骤,很大程度上影响了整个系统的性能。

在自然语言处理领域,常用的中文自动分词方法是机

收稿日期:2005-08-04

**作者简介:**张丽(1980-),女,山东莱芜人,硕士研究生,研究方向为计算机信息控制;黄东,副教授,硕士生导师,研究方向为计算机信息控制、管理信息系统设计与开发。

机械分词。机械分词方法指的是主要依据词典信息,而不使用规则知识和统计信息,按一定的策略将汉字串与词典中的词逐一匹配,如果匹配成功,就加以切分<sup>[1]</sup>,因此需要一个基本词典。常用的有正向最大匹配和逆向最大匹配。

由于在邮件中有些词例如“了”、“的”、“是”等等对分类几乎不起作用,因此把这些词称为停用词。需要一个停用词词典,去掉这些词。

本系统采用正向最大匹配法。中文分词模块直接为后面的特征选择和提取提供原始数据,因此分词效果会直接影响整个系统。

子集。特征提取可以压缩特征空间的维数,剔除干扰特征。目前的特征提取算法主要有以下几种<sup>[3]</sup>:

- \* 文档频次
- \* 相对熵
- \* 互信息
- \* 优势率
- \* 信息增益

本系统中采用信息增益算法。信息增益即 Information Gain,简称 IG,定义如下:

$$IG(t) = - \sum_{i=1}^n P(c_i) \log P(c_i) + P(t) \sum_{i=1}^n P(c_i | t) \log P(c_i | t) + P(\bar{t}) \sum_{i=1}^n P(c_i | \bar{t}) \log P(c_i | \bar{t})$$

IG(t)反映了该词为整个分类所提供的信息量。其中, $P(c_i)$ 表示 $c_i$ 类邮件在语料中出现的概率, $P(t)$ 表示语料中包含特征项 $t$ 的邮件的概率, $P(c_i | t)$ 表示邮件包含特征项 $t$ 时属于类的概率, $P(\bar{t})$ 表示语料中不包含特征项 $t$ 的邮件的概率, $P(c_i | \bar{t})$ 表示邮件不包含特征项 $t$ 时属于类的概率, $n$ 表示类别数。

由上面的公式计算每个特征项的信息增益,进行排序,抽取排在前面一定数量的词作为最终特征集。对于具体多少数目,需要在实验中不断地调整达到最优效果。

经过特征提取,得到两个特征集:垃圾邮件特征集和非垃圾邮件特征集。

### 1.3 邮件特征表示

由前面的两个特征集对邮件进行表示。邮件的表示可采用文档的表示方法。常用的有:布尔逻辑型、向量空间型、概率型和混和型等。向量空间模型(Vector Space Model, VSM)<sup>[4]</sup>是60年代末由Gerard Salton等人提出的。目前已经广泛

应用于文本分类。采用VSM将邮件表示为一 $n$ 维的向量 $(W_1, W_2, \dots, W_n)$ ,其中 $W_k(k=1, 2, \dots, n)$ 表示第 $k$ 个权重, $n$ 为特征集中特征项的数目。

由已得到的特征集和邮件的分词结果,提取邮件的特征项。Winnow算法中邮件权重为布尔值,即 $W_k \in \{0, 1\}$ 。若某个特征词在文本中出现,则其权重为1,否则为0。

### 1.4 分类器的构造

分类器构造是选取一种分类函数,利用邮件样本,训练函数的参数,最终将得到的结果以文件的形式存下来。目前有很多分类函数,包括贝叶斯、K近邻、决策树、支持向量机等。文中采用Winnow算法。

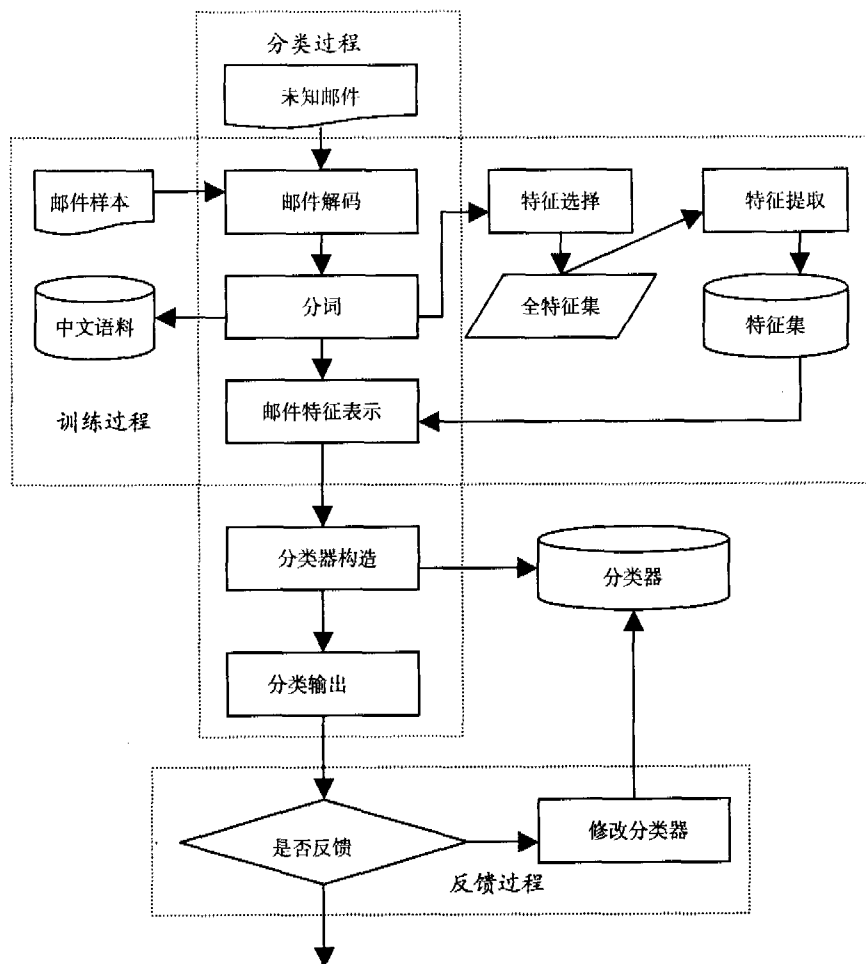


图1 系统流程图

## 1.2 特征选择和提取

### 1.2.1 特征选择

分词后,邮件的内容可以由其中的词来表示。由于样本邮件(垃圾邮件或非垃圾邮件)的属性是可知的。对于垃圾邮件和非垃圾邮件这两个集合中的所有样本邮件,分别将其中的特征合并,可以得到垃圾邮件和非垃圾邮件的全特征集<sup>[2]</sup>。

### 1.2.2 特征提取

全特征集包含了邮件样本出现的所有特征。如果以这样的集合进行下面的分类器构造,不仅效率低,而且由于全特征集包含了很多干扰特征,正确率也不会很高。所以需要选择特征计算公式,对其进行筛选,提取一个特征

### 1.4.1 Winnow 算法介绍

Winnow 算法<sup>[5]</sup>是典型的线性分类算法。对每个类别,训练得到权重向量 $(W_1, W_2, \dots, W_n)$ ,给定一个待分类的邮件 $X = (X_1, X_2, \dots, X_n)$ ,如果 $\sum_{i=1}^n W_i X_i > \theta$  ( $\theta$  为阈值),则将邮件归为垃圾邮件,否则是非垃圾邮件。

### 1.4.2 分类步骤

1) Winnow 算法中初始权重向量为 $(1, 1, \dots, 1)$ ,将训练样本表示为 $X = (X_1, X_2, \dots, X_n)$ ,阈值 $\theta$ 定为训练集中平均每封邮件包含的特征数量。

2) 依次读入训练样本。由于样本的属性是可知的,如果分类错误,可立即调整权重。Winnow 算法是错误驱动的在线学习算法,权重的更新策略如下所示:

\* 如果 $\sum_{i=1}^n W_i X_i > \theta$ ,表示当前分类器预测训练文本属于垃圾邮件,如果实际上邮件不属于垃圾邮件,则降低分类器的权重:对 $i = 1, 2, \dots, n$ ,如果 $X_i \neq 0$ ,则 $W_i = 0$ ;

\* 如果 $\sum_{i=1}^n W_i X_i < \theta$ ,表示当前分类器预测训练文本不属于垃圾邮件,如果实际邮件文本属于垃圾邮件,则提高分类器的权重:对 $i = 1, 2, \dots, n$ ,如果 $X_i \neq 0$ ,则 $W_i = \partial W_i$ ,  $\partial > 1$ 。

### 1.5 未知邮件的分类输出

预处理和训练部分相同,自动分词后用已得到的结果将邮件特征表示。应用已得到的分类器进行分类输出。

## 2 反馈学习

Winnow 是一种在线的错误驱动学习方法,增量式学习非常方便,可以采用和训练阶段完全相同的方法,错误驱动,反馈更新分类器的权重,计算量也很小。由于一般情况下 Winnow 算法中选择的特征都比较多,特征的细小变化对分类效果影响不大,因此可以相隔较长时间才考虑特征重构。

本系统采用在一段时间内不改变特征空间的在线增量式学习方法,隔一个较长的时间后,可以重新学习一次,包括特征选择和各个特征的权重计算。

## 3 实验结果

### 3.1 实验数据

中文垃圾邮件目前还没有像英文垃圾邮件那样,有公共的语料库,目前的语料需要自己收集,这在一定程度上也影响了邮件的分类效果。目前已收集了 500 封垃圾邮件,500 封非垃圾邮件。

### 3.2 评价标准

采用了文本分类的评价标准。设 $N$ 为训练邮件的总数, $A$ 表示实际是垃圾邮件,且分类结果也是垃圾邮件的

数目; $B$ 表示不是垃圾邮件,但分类结果也是垃圾邮件的数目; $C$ 表示是垃圾邮件,但分类结果不是垃圾邮件的数目; $D$ 表示不是垃圾邮件,且分类结果也不是垃圾邮件的数目,则 $N = A + B + C + D$ 。

\* 精度 (Precision)。分类器在一个类别中做出的正确分类与分类器在该类上做出的所有分类的百分比,精度越高表明分类器在该类上出错的概率越小: $P = A / (A + C) \times 100\%$ 。

\* 查全率 (Recall)。分类器在一个类别中做出的确切分类与该类实际应有分类数目的百分比,查全率越高表明分类器在该类上可能漏掉的分类越少: $R = A / (A + B) \times 100\%$ 。

### 3.3 实验结果

由图 2 可以看出(其中横坐标为特征数量),随着特征数量的变化, Precision 和 Recall 都没有很大的变化,而是在一个区间稳定的波动,这说明了 Winnow 算法的稳定。但是同时也可以看出,系统的 Precision 和 Recall 都不是很高,其中最高的才接近 80%。分析原因,一方面是因为中文分词的复杂性,本系统只考虑了机械分词,这必然存在着分词错误的情况;另一方面是因为垃圾邮件的属性是复杂多变的。

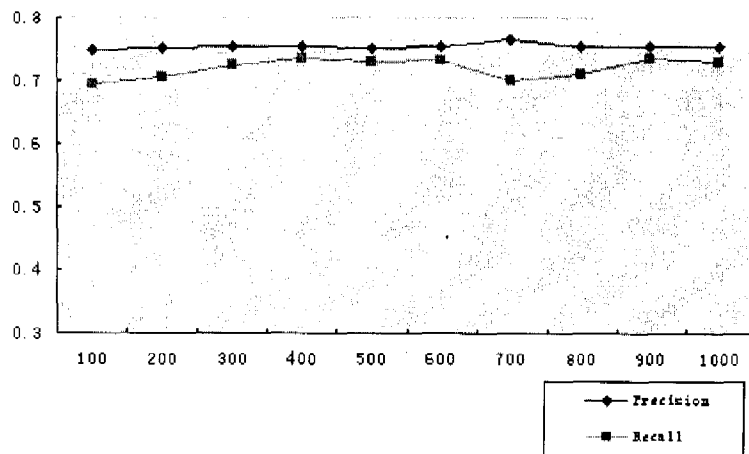


图 2 特征数量对 Precision 和 Recall 的影响

## 4 结束语

文中采用文本分类系统的关键技术,提出了反垃圾邮件引擎的结构模型,同时给出了一些笔者在实现时的具体做法,但是系统的效果还有很多值得提高的地方,将来还需要继续对反邮件引擎系统做进一步研究。

### 参考文献:

- [1] 吴立德. 大规模中文文本处理[M]. 上海: 复旦大学出版社, 1997.
- [2] 姚松源. 文本自动分类系统的研究与实现[D]. 北京: 北京工业大学, 2003.
- [3] 潘文峰. 基于内容的垃圾邮件过滤研究[D]. 北京: 中国科学院研究生院, 2004.

### 3 模型的简化

在一个生产周期中,产品的生产以满足下游生产线的需求为基本要求。因此,主成品的生产根据需求优先安排。假设主成品的需求量远大于其缓冲区的容量,则当主成品缓冲区中产品数量达到容量极限时,上游生产线停止生产主成品,转为生产副成品或同步副成品。待下游生产线将主成品缓冲区中的产品消耗完毕,上游生产线重新切换至主成品的生产。对以上过程进一步分析可以看出,因为上下游之间的主成品缓冲区的大小是固定的,且上下游的生产线对主成品的生产速度也是一定的,所以,在一次主成品和副成品(或同步副成品)的切换中,用于加工副成品(或同步副成品)的时间为下游生产线消耗全部主成品缓冲区里产品的时间,即为定值。因此,在一个生产周期中,主成品的加工时间和切换次数是可以预先确定的,进而确定了主成品的加工数量。基于以上的分析,在计划模型中,就可以不考虑主成品,大大地减少了模型的维数。简化后的目标函数如下:

$$J = J_{\text{主}} + J' \quad (15)$$

$$J' = \min \left\{ \sum_{i=1}^M \sum_{k=1}^N \left\{ \hat{a}_i^T \hat{x}_i(k) + b_i^T [\hat{T}_i \hat{u}_i(k) - \hat{p}_i(k)] + b_i^T [\hat{p}_i(k) - \hat{T}_i \hat{u}_i(k)] + d_i^T \hat{H}_{mi} [\hat{y}_i(k) - \hat{z}_i(k)] + d_i^T \hat{H}_{mi} [\hat{z}_i(k) - \hat{y}_i(k)] + e_i^T \hat{H}_{mi} [\hat{y}_i(k) - \hat{z}_i(k)] + e_i^T \hat{H}_{mi} [\hat{z}_i(k) - \hat{y}_i(k)] \right\} + \sum_{i=1}^M \hat{a}_i^T \hat{x}_i(N+1) \right\} \quad (16)$$

与(1)式相比较,(16)式中的各项参数作以下调整:

① (1)式中的  $x_i(k)$ ,  $u_i(k)$ ,  $\bar{y}_i(k)$ ,  $\bar{z}_i(k)$  基本含义不变,向量维数由考虑主成品的  $n_i + m_i + \bar{m}_i$  维列向量,调整为(16)式中不考虑主成品的  $m_i + \bar{m}_i$  维列向量  $\hat{x}_i(k)$ ,  $\hat{u}_i(k)$ ,  $\hat{y}_i(k)$ ,  $\hat{z}_i(k)$ 。

② (1)式中产品变换矩阵  $H_{mi}$ ,  $\bar{H}_{mi}$  分别调整为(16)式中  $m_i \times (m_i + \bar{m}_i)$  维矩阵  $\hat{H}_{mi}$  和  $\bar{m}_i \times (m_i + \bar{m}_i)$  维矩阵  $\bar{\hat{H}}_{mi}$ 。

③ (16)式中  $\hat{T}_i$  为生产线  $i$  各工位在周期  $k$  加工  $m_i + \bar{m}_i$  种工件所需要的时间,为  $f_i \times (m_i + \bar{m}_i)$  维矩阵。

④ (16)式中  $\hat{p}_i(k)$  为生产线  $i$  各工位在周期  $k$  可用于加工副成品和同步副成品的时间,是从生产周期中扣除主成品生产时间及设备故障维修时间等所剩余的时间,为  $f_i$  维列向量。

约束条件调整:

1) 因为不考虑主成品,则约束条件中去掉了(13)式的下游生产线成品约束,添加时间约束如下:

时间约束:

$$\hat{p}_i(k) = p_i(k) - \bar{p}_i(k)$$

$$\bar{p}_i(k) = [t_{i1}(k) t_{i2}(k) \cdots t_{if_i}(k)]^T$$

$$t_{i1}(k) = t_{i2}(k) = \cdots = t_{if_i}(k) = t_i(k)$$

$$t_i(k) = \frac{B_{mq}}{V_{\max mq} - \bar{V}_{mq}} \times N$$

$$V_{\max mq} = \max \{ V_{i mq} \}$$

$$i = 1, 2, \dots, M; k = 1, 2, \dots, N$$

式中,  $\bar{p}_i(k)$  为生产线  $i$  在周期  $k$  用于加工主成品的加工时间,是  $f_i$  维列向量;  $B_{mq}$  为生产线  $i$  在周期  $k$  所生产的第  $q$  种主成品的下游输入缓冲区大小;  $V_{i mq}$  为生产线在周期  $k$  生产第  $q$  种主成品的速度;  $V_{\max mq}$  为  $M$  条生产线中生产第  $q$  种主成品的最大速度;  $\bar{V}_{mq}$  为下游生产线对第  $q$  种主成品的加工速度;  $\bar{N}$  为周期  $k$  内主产品和副产品(或同步副产品)的切换次数。

2) 其余约束条件与约束(2)~(12)式、(14)式相似,条件中的向量或矩阵的维数减去主成品的种类数  $n_i$ 。

### 4 实例分析

设上游生产线有4条,每条生产线上各有3个工位,即  $f_i = 3$ ,生产周期  $N = 10$ ,各生产线加工的主成品、副成品和同步副成品的种类分别为(2,1,1),(2,1,1),(2,1,1),(1,1,1)。用(1)式所建模型的约束条件个数为770个,变量个数为960个。采用(16)式简化后,模型的约束条件个数为490个,变量个数为680个。目标函数的维数被大大地减少,说明这种简化方法是有效的。

### 5 结束语

根据 open-shop 型生产线不同产品工序可不同的特性,建立了一种多 open-shop 生产线的协调生产计划模型。同时,从模型建立的角度提出了一种模型简化的方法,为下一步采用合适的算法,取得快速、精确的最优值创造了有利条件。该模型的算例仿真将另文给出。

### 参考文献:

- [1] 王廷平.多生产线协调生产计划的研究[D].南京:东南大学,2004.
- [2] 蒋珉,王廷平.基于关联加权预测的多生产线协调生产计划的研究[J].东南大学学报,2004,34(增刊):174-181.
- [3] 丁广太,涂奉生.串行生产线的状态及输出与其缓冲区容量的关系分析[J].南开大学学报,1999,32(4):107-114.
- [4] 严洪森.一种基于关联预测的车间生产计划的最优分解方法[J].系统工程理论与实践,1997,17(12):58-63.
- [5] 张燕红,蒋珉.含有限缓冲区的上游生产线协调生产调度[J].微机发展,2004,14(10):117-119.

(上接第172页)

- [4] 张东礼,汪东升,郑纬民:基于VSM的中文文本分类系统的设计与实现[J].清华大学学报(自然科学版),2003,43(9):1288-1291.

- [5] Littlestone N. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm[J]. Machine Learning, 1988, 2(4): 285-318.