中文垃圾邮件过滤系统中的特征提取算法①

白飞云，王新房

(西安理工大学 自动化与信息工程学院，西安 710048)

摘 要：针对垃圾邮件过滤，首先对获取的垃圾邮件及合法邮件进行分词，预处理，构建文本矢量，然后用四种常用的特征词提取方法进行矢量降维，再在此基础上，给出了一种综合性的特征词提取算法，即按照各个评估函数的排序结果，取它们交集的前 n 个特征词作为候选词进行分类测试,仿真比较了各个算法中 n 对分类结果的影响，从而验证了该算法的有效性。

关键词：垃圾邮件过滤；邮件预处理；特征提取；Rocchio 方法；评价指标

Feature Selection Method in Chinese Spam Filtering

BAI Fei-Yun, WANG Xin-Fang

(School of Automation& Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: The paper, aimmed at spam filter, at first separationing, preproccessing and building text vector for the obtained spam mails and legitimate mails, then proccessing vector dimensional reduction using four common key extraction methods, and based on this, presents a comprehensive key extraction algorithm, which takes front n key words of their intersection as a candidate word for classification test according to sort results of each assessment function. Finally, Simulation verifies the effection of "n" on the classification in the algorithm, thus verifying the effectiveness of the proposed algorithm.

Key words: spam filtering; preprocessing mail; feature selection method; rocchio; evaluation indicator

随着国际互联网 Internet 的发展和普及，电子邮件以其方便、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，就包括垃圾邮件[1]。由于垃圾邮件的泛滥，引起了广大研究者的极大关注，从而提出了关于垃圾邮件问题的多种解决方法。本文主要讨论基于内容的垃圾邮件过滤，这种过滤技术最终归结为文本分类的问题，文本分类就存在特征词集庞大的弊端，进行分类时不可能把所有的词都选为特征词，一方面参杂的干扰词会降低分类精度；另一方面这样的高维矢量大大的增加了计算的复杂度，浪费计算机资源。

因此，要分类首先要进行特征词选择。目前常用的特征词选择算法包括信息增益，互信息，文档频率等，各种算法各有利弊，但都是基于词频统计的基础上进行的，可以考虑将其综合达到更好的提取效果，从而提高分类指标。

1 邮件预处理

与英文邮件不同，中文邮件中词与词之间的界限也不像英文那样明显，所以必须对邮件正文进行分词，本文采用中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS (Institute of Computing Technology Chinese Lexical Analysis System)。预处理包括以下几步：

1) 垃圾邮件为了干扰垃圾邮件过滤器的识别，会有意的在邮件中插入许多无关、无用的字符和标记，如"法*轮功"，干扰分词结果，因此在分词对原分词算

① 收稿时间:2011-07-05;收到修改稿时间:2011-10-23

法更改删除这些干扰。

2) 分词后含有大量的口语化常用词，如一些语气助词，人称代词等，通过设置停用词表，删除这些字词。

3) 本文针对的是中文垃圾邮件的过滤，所以对于分词后的英文字符串不作保留，仅对网址，￥，$类有特殊意义的字符串进行保留。

2 特征词提取及权值计算

特征提取是指从分词产生的特征词集中选取一部分能反映其内容信息的词语及其权重的计算。邮件分词处理后，形成了一个表征文本类的原始特征词集 $\mathrm{T}=\{\mathrm{t}_1,\mathrm{t}_2,\cdots,\mathrm{t}_n\}$，其中 n 为特征词集总量，这一特征集的维数一般都比较大。因此如何选择特征词，如何降维就成了研究的重点和难点。

最常用的方法就是采用某种评估函数对每一个特征词进行计算,然后按照计算结果的高低排列,数值大于预先设定的阈值的特征词被选取。评估函数有文档频次，信息增益，互信息，类间信息差。

2.1 常用的特征词提取方法

1) 频度/逆文档频度法(DF)，它认为特征在文本集中出现的次数越多，对文本分类的贡献越大[2]:

$$DF(w)=\text{训练集中出现}w\text{的文本数} \tag{1}$$

2) 信息增益(IG),它采用统计某个特征项在一篇文档中出现或不出现的次数来预测文档的类别：

$$\begin{aligned} IG(w)=&-\sum_{i=1}^{m}P(c_i)\lg P(c_i)\\ &+P(w)\sum_{i=1}^{m}P(c_i|w)\lg P(c_i|w)\\ &+P(\overline{w})\sum_{i=1}^{m}P(c_i|\overline{w})\lg P(c_i|\overline{w}) \end{aligned} \tag{2}$$

其中，$P(c_i)$表示类文档在语料中出现的频率；$P(w)$表示语料中包含特征项的文档频率；$P(c_i|w)$表示文档包含特征项时属于类的条件概率；$P(\overline{w})$表示语料中不包含特征项的 w 文档频率；$P(c_i|\overline{w})$表示文档不包含特征项 w 时属于 c_i 类的条件概率[2]。

3) 互信息(MI)是用来衡量两个随机变量的依赖程度的：

$$MI(w,c)=\log\frac{A\times N}{(A+C)\times(A+B)} \tag{3}$$

N 表示训练语料中的文档总数；A 表示包含特征项 W 且属于类别 c 的文档频率；B 表示包含特征项 W 但不属于类别 c 的文档频率；C 表示不包含特征项 W 但属于类别 c 的文档频率[2]。

4) 类间信息差(D), 考虑到如果一个词在每个类中出现的概率值都比较接近，那么这个词在分类过程中的作用就不大，因为这些词语的出现对判断文档的类别没有多大的作用。

$$P_i(w)\approx f_i(w)=\frac{num_i(w)}{|C_i|} \tag{4}$$

$num_i(w)$为在类别 C_i 中出现 w 的文档数；$|C_i|$ 为类别 C_i 中出现的文档总数。

$$D(w)=f_{\max}(w)-f_{\min}(w) \tag{5}$$

$f_{\max}(w)$和 $f_{\min}(w)$ 分别为特征词 w 的最大频率词和最小频率词[3]。

2.2 特征权重的计算

将邮件表示为文本矢量模型，这种模型中最常用的权重就是 TF-IDF 权重函数。

$$w(t,d)=\frac{tf(t,d)\times\log\left(\frac{N}{nt}+0.01\right)}{\sqrt{\sum_{t\in d}\left[f(t,d)\times\log(N/nt+0.01)\right]^2}} \tag{6}$$

其中，$w(t,d)$为词 t 在文本 w 中的权重；$tf(t,d)$为词 t 在文本中的词频；N 为训练文本的总数；nt 为文本集出现 w 的文本数。

3 特征词选择方法

在特征提取中，针对各个评估函数都是依据特征词出现的概率或者频率来计算的。单个采用某个评估函数必然存在优缺点，考虑一个综合方案能结合各个评估函数，达到更好的降维效果。算法步骤如下：

1) 针对邮件这种两类分类问题，先分别用评估函数对特征词进行权值计算，对每个评估函数进行类间排序。

2) 按照各个评估函数的排序结果，取它们的交集的前 n 个特征词作为候选特征词。

3) 以这 n 个特征词构成特征向量 $W=\{w_1,w_2,w_3,\cdots,w_n\}$，用 TF-IDF 法给邮件矢量赋权值。

4) 选择一个分类算法，通过仿真研究分类效果与特征词数目 n 的关系以便取得一个合适的 n 。

4 仿真分析

4.1 分类方法及评价指标

分类算法采用 Rocchio 方法(相似度计算法)，根据待分类邮件向量离正常邮件和垃圾邮件的中心向量的距离来判断文档的类别属性[4]。首先根据算术平均为每类文本集生成一个代表该类的中心向量，然后计算测试邮件集每封邮件的向量与两个中心向量的距离（余弦距离），也就是相似度,将新邮件分到相似度大的那一类。这种算法简单，能够满足邮件过滤中对实时性的要求。

评价指标根据垃圾邮件的常用评价指标，假设待测试的邮件集合中共有 N 封邮件：

表 1 垃圾邮件系统判定情况分布(单位：封)

	实际为垃圾邮件	实际为合法邮件
判为垃圾邮件	A	B
判为合法邮件	C	D

定义如下的几个评价指标来衡量不同垃圾邮件过滤系统的性能：

a) 召回率（Recall）：$R=\frac{A}{A+C}$，及垃圾邮件检出率。这个指标反映了过滤系统发现垃圾邮件的能力，召回率越高，“漏网”的垃圾邮件就越少。

b) 正确率（Precision）：$P=\frac{A}{A+B}$，垃圾邮件检对率。正确率反应了过滤系统“找对”垃圾邮件的能力，正确率越大，将合法邮件误判为垃圾邮件的可能性越小。

c) 精确率(Precision)：$Accur=\frac{A+D}{N}$，即对所有邮件(包括垃圾邮件和合法邮件)的判对率。

d) F 值：$F=\frac{2PR}{R+P}$，F 实际上是召回率和正确率的调和平均，它将召回率和正确率综合成一个指标[5]。

4.2 仿真结果及分析

本文语料库是在中文自然语言处理开放平台（CNLP platform）下载的，由中科院计算所郎皓，王小冷，王斌搜集，包括垃圾邮件合法邮件 2000 个。(包括原文（HTML）格式，文本格式，分词后格式（但未校对））。本文采用的是文本格式。

测试时，在其他条件不变的前提下，将语料分成三份，分别取其中一份做测试集，其他两份做训练集，进行分类，最后取三次的平均值作为测试结果。经过方针测试，各个性能指标与特征词数目的关系如下图所示：

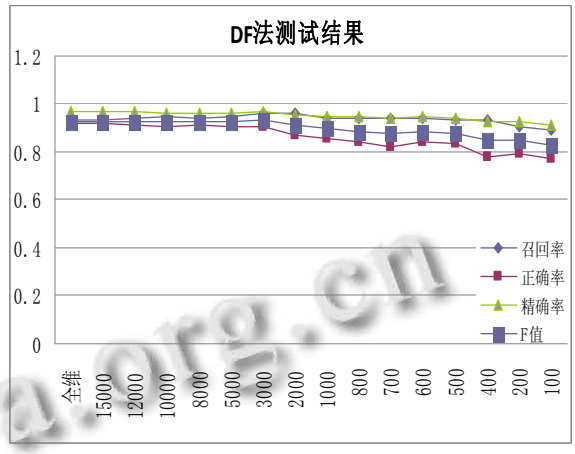

图 1(a)　DF 算法随维数 n 变化测试结图

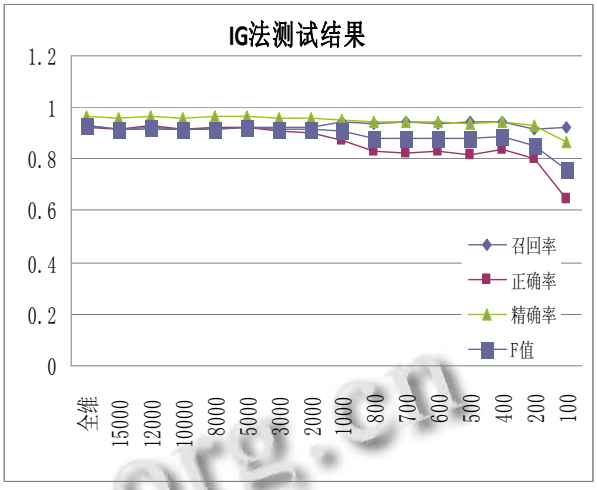

图 1(b)　IG 算法随维数 n 变化测试结果图

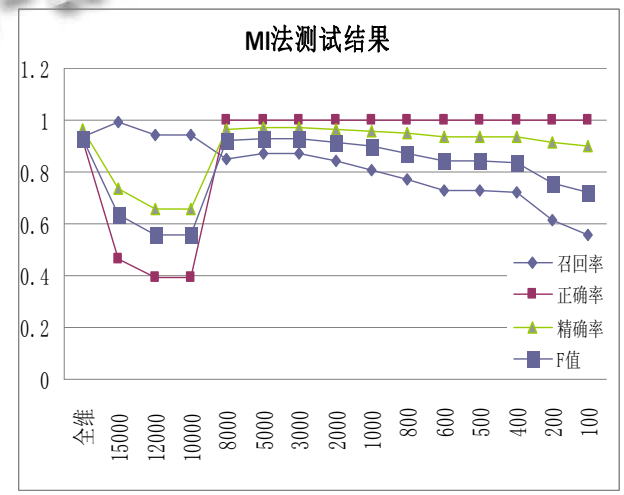

图 1(c)　MI 算法随维数 n 变化测试结果图

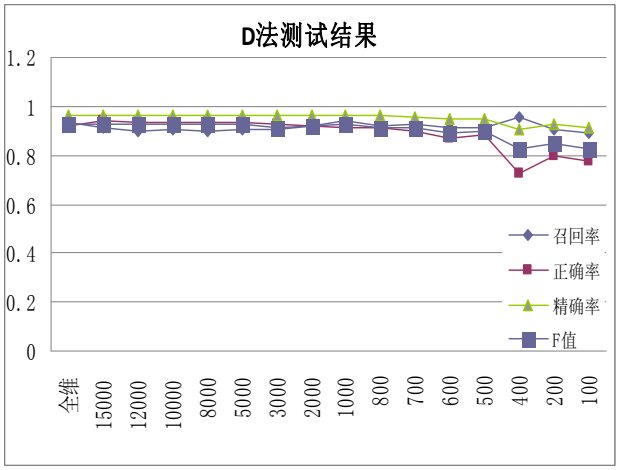

图 1(d) D 算法随维数 n 变化测试结果图

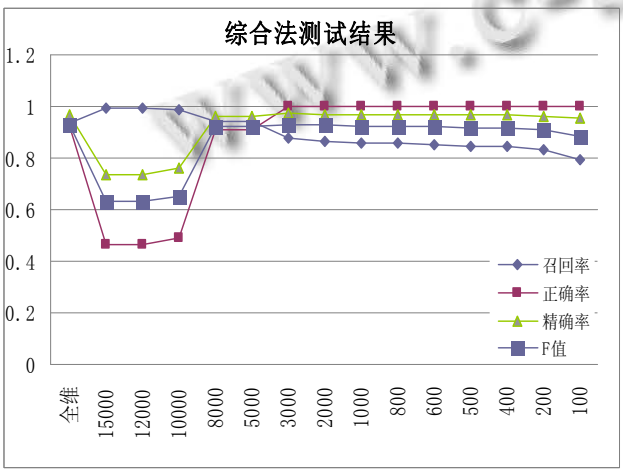

图 1(e) 综合算法随维数 n 变化测试结果图

从上述表中可以看出，在 200 维的时候，各个算法的评价指标有下降的趋势，MI 算法受维数的影响最大，曲线波动比较明显。而综合法曲线较平稳，各项参数始终保持在 0.8 以上。

提取以上各个算法在 n=200，n=100 维时的数据如下表：

表 2 各种特征提取算法 n=200 维测试数据

n	R	P	Accur	F 值	方法
200	0.913	0.801	0.928	0.854	IG
200	0.613	1	0.911	0.76	MI
200	0.907	0.782	0.92	0.836	DF
200	0.906	0.795	0.924	0.847	D
200	0.833	1	0.961	0.829	综合

表 3 各种特征提取算法 n=100 维测试数据

n	R	P	Accur	F 值	方法
100	0.92	0.649	0.866	0.76	IG
100	0.56	1	0.898	0.718	MI
100	0.867	0.764	0.907	0.812	DF
100	0.893	0.774	0.915	0.829	D
100	0.793	1	0.952	0.885	综合

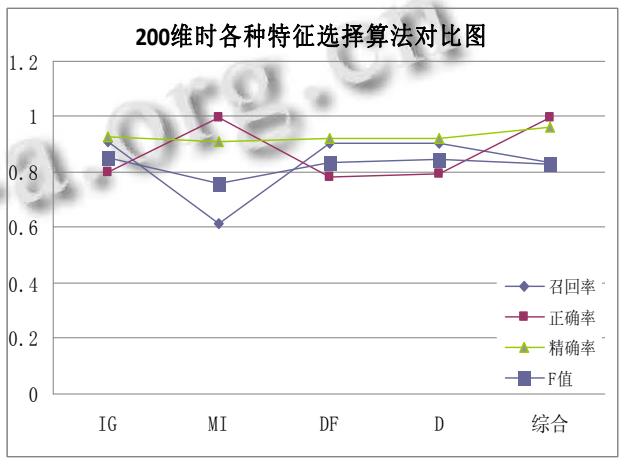

图 2 n=200 时各个算法评价指标对比图

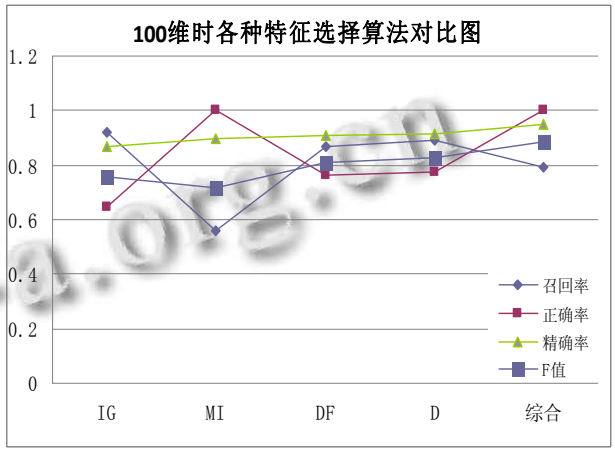

图 3 n=100 时各个算法评价指标对比图

5 实验结果分析

针对以上测试数据得到如下 4 点结论：

1) 分析 MI 算法在高维时的拐点，是因为 MI 算法得分受词条的边缘概率影响，对于有相等条件概率的一些特征词，低频词比常用词的得分还要高，因此对于频率相差很大的特征词，得分是不具备可比性的[4]。从下图（图 4）可以看出：10000 维左右权值得分相差较大，出现了边缘概率影响。

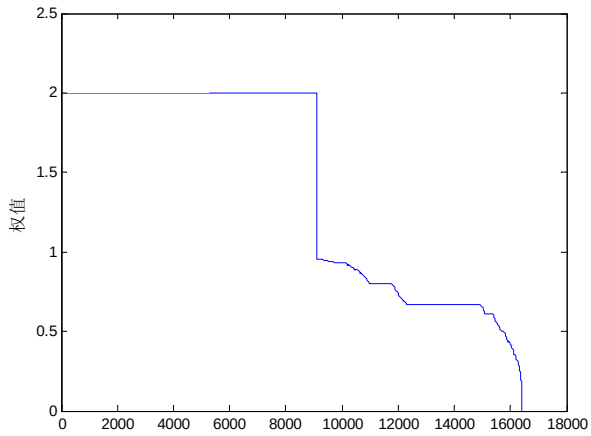

图 4　MI 算法权值分布图

2) 维数不是越高越好，各种算法 5000 维和全维的各项指标几乎相近，低维时 D 算法,DF 算法及综合算法相对有较好的综合特性,MI 算法正确率较高，IG 算法召回率较高。

3) 分析 200 及 100 维的数据，从召回率及正确率这两个评价指标发现：MI 算法容易漏判垃圾邮件，而 IG 算法又容易过滤过度，使用户损失正常邮件。综合方法能够充分利用四种算法的优点，更好的平衡垃圾邮件的检出率和检对率这两个指标。

4) 分析原因，DF 算法保证了取到的特征词是高频词汇；D 算法从类间信息的角度考虑，去除了在两类中概率相近的词语，这就包括常用词，罕见词。IG 算法考虑了单词未出现的情况；而 MI 算法则能够帮助取到一些区分度高的低频特性。取每类的公共集就可以使取到的特征词兼顾到以上各个算法的优点。

6　结语

垃圾邮件用文本矢量表示的最明显的特点就是高维性，针对常用特征提取算法各有所长的特点，本文对各种特征提取算法进行仿真，在分析各种特征提取算法的优缺点的基础上，用一种综合性的特征提取算法，即提取特征词时同时考虑它在每个特征提取算法中的权值排序，取其公共词集构成特征矢量，经分类测试得到了满意的效果。一方面考虑到语料集的限制，对测试肯定会造成一定程度的影响，可以通过加大语料库来提高分类的精度，另一方面 Rocchio 方法本身简单但是受训练文档中的噪声影响较大，同时对某些特征空间非线性可分情况没有处理能力[4]。如果使用 winnow 分类器这类有自修正能力的分类器，相信可以在此基础上进一步提高过滤效果。

参考文献

1 曹麒麟,张千里.垃圾邮件与反垃圾邮件技术.北京:人民邮电出版社,2003.34–56.

2 侯汉清.文本自动标引与自动分类研究.南京:东南大学出版社,2009.57–64.

3 谷波,刘开瑛.中文文本分类中一种简单高效的特征词选择方法.计算机研究与发展,2005,42(增刊):359–360.

4 戴文华.基于遗传算法的文本分类及聚类研究.北京:科学出版社,2008.46–47.

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

6 李晓飞.垃圾邮件过滤算法研究及系统实现.南京:南京理工大学,2008.

（上接第 205 页）

4 Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowledge and Data Engineering, 2005,17(6):734–749.

5 Wei CP, Yang CS, Hsiao HW. A collaborative filtering-based approach to personalized document clustering. Decision Support Systems, 2008,45(3):413–428.

6 Liang HZ, Xu Y, Li YF, et al. Connecting Users and Items with Weighted Tags for Personalized Item Recommendations. Proc. of the 21st ACM Conference on Hypertext and Hypermedia, 2010: 51–60.

7 Zheng N, Li QD. A recommender system based on tag and time information for social tagging systems. Expert Systems with Applications,2011,38(4):4575–4587.

8 Kim HN, Alkhaldi A, Saddik AE, et al. Collaborative user modeling with usergenerated tags for social recommender systems. Expert Systems with Applications, 2011,38(7): 8488–8496.

9 Su JH, Wang BW, Hsiao CY, et al. Personalized rough-set-based recommendation by integrating multiple contents and collaborative information. Information Science, 2010,180(1): 113–131.