一种垃圾邮件协作过滤模型

章雅娟, 张 虹

(西南大学计算机与信息科学学院, 重庆 400715)

[摘要] 传统垃圾邮件过滤方法大多基于单一技术, 已不能有效阻止不断出现的新型垃圾邮件. 在分析传统单一过滤方法的基础上, 本文提出了一种垃圾邮件协作过滤模型 SCF, 它可以弥补单一过滤技术的缺点, 发挥各技术的优势, 从而有效地过滤垃圾邮件. 实验结果表明, 该模型具有较高的垃圾邮件过滤召回率和正确率.

[关键词] 垃圾邮件, 反垃圾邮件, 协作过滤

[中图分类号] TP393. 0 [文献标识码] A [文章编号] 1001-4616(2010) 04-0139-05

A Spam Collaborative Filtering Model

Zhang Yajuan, Zhang Hong

(Faculty of Computer and Information Science & Software, Southwest University, Chongqing 400715, China)

Abstract Most traditional spam filtering methods are based on single technology, so they can't effectively prevent the emerging new type of spams. After we have analyzed traditional filtering methods based on single technology, a spam collaborative filtering model SCF is proposed. The SCF model can make up the shortcomings and play the advantages of each technology. The experimental results show that this newly developed SCF model has better recall and precision in filtering spams.

Key words spam, anti-spam, collaborative filter

在网络使用日益频繁的今天, 电子邮件逐渐成为人们相互交流的主要工具. 但是, 越来越多的垃圾邮件影响了正常邮件的传输. 对于特征不断变异的垃圾邮件, 传统的反垃圾邮件方法显得力不从心. 从单项、单点的技术研究转移到对多技术体系融合、协作式垃圾邮件技术体系的研究已成为一种趋势[1, 2].

目前针对垃圾邮件的对策, 从技术层面上讲, 主要包括 3类技术, 一是基于 IP、域名和路由等的过滤技术; 二是基于内容的过滤技术; 三是基于行为的过滤技术. 在实际应用中, 对邮件内容进行过滤被认为是目前最有效的垃圾邮件过滤方法, 它主要包括朴素贝叶斯、K 邻近、支持向量机和神经网络等文本分类方法.

朴素贝叶斯算法[3]性能较好, 被广泛地使用, 但该算法成立的前提是各属性之间相互独立, 当数据集不满足这种独立性假设时, 分类的准确度较低[4]; K 邻近方法[2]通过计算文本之间的相似度以进行文本分类, 当数据存在噪声或不相关属性时, 该方法的准确率可能会受影响, 且其在分类阶段所需的时间较长; 支持向量机[5]方法, 通过构造最优线性分类面来指导分类, 它在解决小样本学习、非线性及高维模式识别问题中表现较好, 但是不太容易拟合, 而且训练和分类速度较慢; 神经网络是一种智能的文本分类算法, 它所需要的规则是在网络构造过程中自动隐式建立的, 无需通过人工归纳整理, 然而, 传统的神经网络算法学习收敛速度慢、容易陷入局部极小值且受训练样本影响大[6]; 基于人工免疫[7]的邮件过滤技术借鉴了生物免疫系统抗体对抗原的识别原理, 具有免疫机制的自学习、自适应以及鲁棒性等优点, 但该算法训练时间长, 检测性能稳定性较差.

从上述分析可知, 单一过滤技术在垃圾邮件过滤中存在各自的缺点, 过滤性能也不能达到理想的效

收稿日期: 2010-07-10

基金项目: 中央高校基本科研业务费专项资金 (XDJK2009C018).

通讯联系人: 张 虹, 副教授, 研究方向: 人工智能、视觉认知计算. E-mail: zhangh@swu.edu.cn

果.因此,为了弥补单一过滤技术的缺陷,本文提出了一种垃圾邮件协作过滤模型——SCF模型(Spam Collaborative Filtering Model).

1　SCF模型设计

SCF模型的概念化设计如图 1所示,主要由预处理模块、训练模块、学习模块以及协作过滤模块等组成.下面分别对这几个主要模块进行说明.

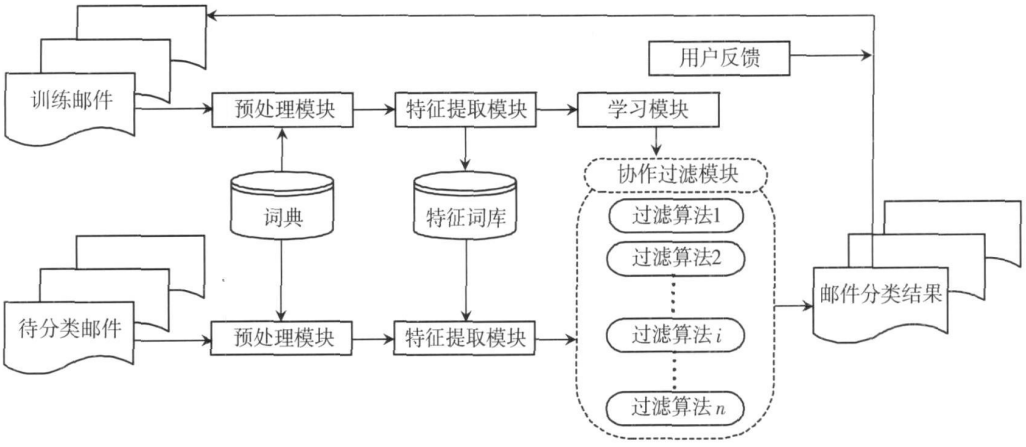

图 1　SCF 模型

Fig.1　Spam collaborative filtering model

(1) 预处理模块:为了实现邮件内容的过滤,首先要对其进行分词操作,得到邮件文本的特征向量.本文采用 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)[8]分词系统,该系统的分词正确率高达 97.58%.

(2) 特征提取模块:分词后会产生一个庞大的词集,但并不是所有词对分类来说都是必要的,而且过高维数的运算,也会导致过高的时空复杂度.

(3) 学习模块:由协作过滤模块的构造和增量学习两个子模块组成.具体步骤包括:通过学习训练邮件集中的邮件样本,构造出协作过滤模块;增量学习模块是指,用户对邮件分类结果进行反馈,然后将这些邮件作为新的训练集对协作过滤模块进行训练学习.

(4) 协作过滤模块:将待分类邮件输入已构造好的协作过滤模块,得到过滤效果.在过滤一定数量的邮件后,结合用户的反馈对邮件进行分类标记,然后将其输入学习模块进行反馈型学习.

2　协作过滤算法

单一邮件过滤技术有很大的局限性,在过滤上无法达到理想的效果.针对本文 SCF 模型的协作过滤模块,通过结合不同的垃圾邮件过滤技术,设计了两种协作过滤算法:层次过滤和组合过滤.

2.1　层次过滤

2.1.1　层次过滤流程

层次过滤算法的流程如图 2所示,它主要将多个过滤算法分布在整个过滤模块的不同层次上,以对邮件进行过滤.

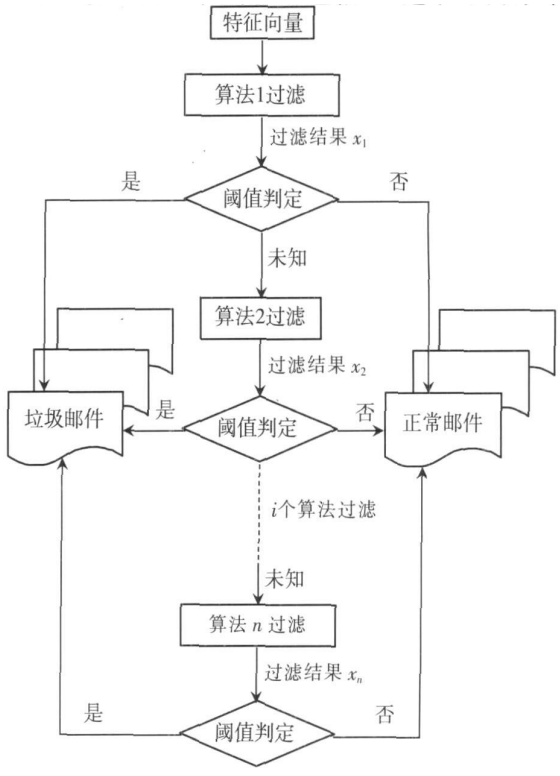

图 2　层次过滤算法流程图

Fig.2　Flow chart of hierarchical filtering algorithm

2.1.2 层次过滤算法

由图 2可知, 层次过滤对待分类邮件进行一层一层的筛选, 由过滤结果判定邮件的分类结果或者将其输入下一层进行过滤, 具体算法步骤如下所示:

```
//共有 n种过滤算法, 算法 i对待分类邮件 x的过滤结果为 x_i
输入: 待分类邮件的特征向量
输出: 待分类邮件所属类别, 垃圾邮件或正常邮件
算法步骤:
    for( int i= 1; i < n; i++ )
        将特征向量输入算法 i进行过滤操作, 得到过滤结果 x_i;
        if(过滤结果 x_i > 阈值 θ_i1 )
            return该邮件为垃圾邮件;
        else if(过滤结果 x_i < 阈值 θ_i2 )
            return该邮件为正常邮件;
    将特征向量输入算法 n进行过滤操作, 得到过滤结果 x_n;
    if(过滤结果 x_n >= 阈值 θ_n1 )
        return该邮件为垃圾邮件;
    else return该邮件为正常邮件;
```

2.2 组合过滤

组合过滤的流程图如图 3所示, 它主要将待分类邮件 x的特征向量输入 n个过滤算法, 对它们的过滤结果进行组合.

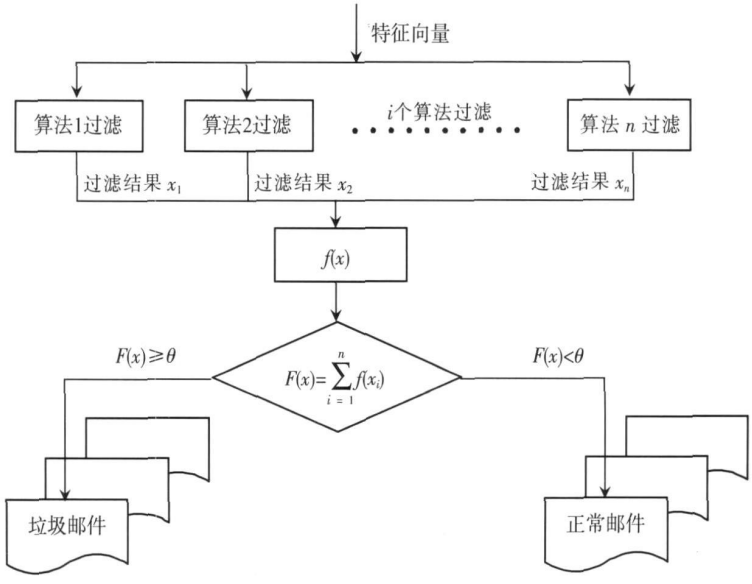

图 3 组合过滤算法流程图

Fig.3 Flow chart of combined filtering algorithm

假设新邮件 x转化成的特征向量经算法 i过滤的结果为 x_i, 由于邮件的类别只有垃圾邮件和正常邮件两类, 所以使用函数 $f(x)$将过滤的结果映射为 0-正常邮件或 1-垃圾邮件:

$$f(x_i)=\begin{cases}1, & x_i \geqslant \theta_i\\ 0, & x_i < \theta_i\end{cases},\quad \theta_i \text{为算法 } i\text{的判定阈值.} \tag{1}$$

最终, 整个混合模型对于邮件 x的过滤结果表示为:

$$F(x)=\sum_{i=1}^{n} f(x_i)\,. \tag{2}$$

上式中 n为过滤算法个数, 若该模型对邮件 x的过滤结果大于或等于某一阈值 $F(x)\geqslant\theta$, 则判定该邮件 x为垃圾邮件, 否则判定其为正常邮件.

3　实验与分析

3.1　数据集

实验采用 CCERT提供的中文邮件数据集 CDSCE(CCERT Data Sets of Chinese Emails)[9]作为语料库. 该论文在训练阶段采用的是 CCERT提供的中文邮件语料库中 2005年 6月份的邮件. 而测试阶段采用的是 2005年 7月份的邮件. 6月份的语料库包含 9272封正常邮件和 25 088封垃圾邮件; 7月份共有垃圾邮件 20 308封, 正常邮件 9042封, 各占总邮件量的 69% 和 31%.

3.2　性能评价指标

假设待测试的邮件共有 N 封, 垃圾邮件过滤系统的判定结果如表 1所示.

表 1　垃圾邮件过滤系统判定情况分布 /封

Table 1　Distribution of spam filtering system results

	垃圾邮件	正常邮件
系统判定为垃圾邮件	A	B
系统判定为正常邮件	C	D

下面定义两个技术评价指标:

召回率(Recall): $\mathrm{Recall}=\frac{A}{A+C}\times 100\%$, 即垃圾邮件检出率. 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"的垃圾邮件越少.

正确率(Precision): $\mathrm{Precision}=\frac{A}{A+B}\times 100\%$, 即垃圾邮件检对率. 这个指标反映了过滤系统"找对"垃圾邮件的能力, 正确率越大, 将非垃圾邮件误判为垃圾邮件的数量越少.

3.3　实验与结果分析

在基于内容的过滤技术中, 朴素贝叶斯具有较好的过滤性能, 而 BP神经网络与人工免疫具有较强的学习能力. 因此, 本文选取这 3种垃圾邮件过滤技术, 分别采用层次过滤算法和组合过滤算法对 SCF模型进行测试.

3.3.1　协作过滤算法参数测试

从 2005年 7月的数据集中随机抽取正文内容不同的垃圾邮件和正常邮件各 1 000封作为测试集, 分别对两种协作过滤算法调整相关参数进行性能测试.

3.3.1.1　层次过滤

由于各单一算法对垃圾邮件的过滤性能不同, 所以层次过滤算法中各层所选的过滤算法将影响整个 SCF模型的性能. 分别调整朴素贝叶斯算法、BP神经网络算法和人工免疫算法所应用的层次, 及其各自所选取的判定阈值, 以对 SCF模型进行性能测试. 由实验可得, 如下参数设置能达到较好的垃圾邮件过滤性能:

层次 1算法——BP神经网络算法, $\theta_{11}=0.99$; $\theta_{12}=0.01$;

层次 2算法——人工免疫算法, $\theta_{21}=0.6$;

层次 3算法——贝叶斯算法, $\theta_3=0.8$

SCF模型性能: 召回率 = 92.20%, 正确率 = 96.24%.

表 2　层次过滤判定情况分布 /封

Table 2　Distribution of hierarchical filtering results

	垃圾邮件误判数	正常邮件误判数	垃圾邮件识别数	正常邮件识别数
层次 1	1	1	271	176
层次 2	0	4	435	0
层次 3	77	31	216	788
混合算法	78	36	922	964

3.3.1.2　组合过滤

组合过滤算法中阈值的选取至关重要, 直接影响着该 SCF模型的性能. 函数 $f(x)$中主要针对不同的过滤结果, 选取不同的阈值, 从而将其映射成 1或 0. 分别调整朴素贝叶斯、BP神经网络和人工免疫所选取的阈值, 以使 SCF模型达到较好的性能. 由实验可得, 如下参数设置能达到较好的垃圾邮件过滤性能:

算法 1——BP神经网络算法, $\theta_1=0.4$;

算法 2——人工免疫算法, $\theta_2=0.5$;

算法 3——贝叶斯算法, $\theta_3=0.3$

SCF模型性能: 召回率 = 94.90%, 正确率 = 95.28%.

3.3.2　综合测试

从 2005年 7月的数据集中随机抽取垃圾邮件和正常邮件各 9 000封, 即 18 000封作为测试样本邮件. 将这些邮件分为 10份, 每份 1 800封, 每次取一定份数作为测试集, 分别对贝叶斯算法、BP神经网络算法、人工免疫算法、层次过滤算法以及组合过滤算法进行性能测试并比较, 结果如表 3和表 4所示.

表 3 召回率比较

Table 3 Recall comparison

邮件样本数	贝叶斯/%	BP神经网络/%	人工免疫/%	层次过滤/%	组合过滤/%
3600	97.72	98.27	94.00	97.94	98.39
7200	97.25	98.14	96.00	98.03	98.61
10800	96.87	97.74	96.59	97.76	98.26
14400	96.51	97.75	96.86	97.63	98.08
18000	96.68	97.98	97.29	97.90	98.33

表 4 正确率比较

Table 4 Precision comparison

邮件样本数	贝叶斯/%	BP神经网络/%	人工免疫/%	层次过滤/%	组合过滤/%
3600	97.94	72.85	97.97	98.16	97.30
7200	97.84	72.67	97.60	98.00	97.10
10800	97.25	73.53	97.24	97.47	96.46
14400	96.86	72.83	97.00	97.23	95.95
18000	96.91	72.86	97.02	97.26	96.02

从表中可以看出, 应用较广泛的贝叶斯算法, 在实验中有较好的过滤效果, 召回率和准确率都保持在96%以上, 但由于该算法不具备反馈学习能力, 所以随着测试集的增加, 它的召回率有所下降. SCF模型不论采用层次过滤还是组合过滤, 其召回率始终高于贝叶斯算法, 对垃圾邮件有较好的识别能力. 其中, 采用组合过滤算法时, SCF模型的召回率始终保持在 98%以上; 而采用层次过滤算法时, SCF模型不论在召回率还是正确率上都明显优于单一的贝叶斯算法.

当测试集较小时, SCF模型的优势不是很明显, 随着测试集的增加, 其性能会慢慢的提高, 整体效果都高于单一过滤算法. 这是由于 SCF模型不仅具有贝叶斯过滤算法较高的过滤能力, 同时还具有人工免疫算法和人工神经网络算法的学习更新能力, 具有极强的动态特性.

4 结论

本文研究了将多技术结合的垃圾邮件过滤技术, 设计了一种协作过滤模型 SCF, 并选用贝叶斯算法、BP神经网络算法和人工免疫算法对该模型进行测试和对比. 实验证明, SCF模型对垃圾邮件的过滤性能要高于单一过滤技术的性能, 具有较高的召回率和准确率. 下一步的研究工作可着眼于 SCF模型中的协作过滤算法的改进, 通过对各过滤算法设置权重或应用其他融合技术, 以提高 SCF模型的过滤性能.

[参考文献]

[1] Gu-Hsin Lai, Chia-Mei Chen, Chi-Sung Laih, et al. A collaborative anti-spam system[J]. Expert Systems With Applications, 2009, 36(3): 6645-6653.

[2] Mehmet Aci, Cigdem Inan, Mutlu Avci. A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm[J]. Expert Systems With Applications, 2010, 37(7): 5061-5067.

[3] Muhammad N Marsono, M Watheq El-Kharashi, Fayez Gebali. Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification[J]. Computer Networks, 2009, 53(6): 835-848.

[4] 孔维华, 刘继承, 陈娟. 基于优化 Naïve Bayes的垃圾邮件过滤[J]. 计算机安全, 2009(1): 18-20.

[5] Yu Bo, Xu Zongben. A comparative study for content-based dynamic spam classification using four machine learning algorithms[J]. Knowledge-Based Systems, 2008, 21(4): 355-362.

[6] 郭守团, 徐志根. 基于 BP神经网络的垃圾邮件过滤器研究[J]. 计算机安全, 2009, 12: 19-20.

[7] Guzella T S, Mota-Santos T A, Uchôa J Q, et al. Identification of SPAM messages using an approach inspired on the immune system[J]. Biosystems, 2008, 92(3): 215-225.

[8] 中国科学院计算所. 汉语词法分析系统 ICTCLAS[CP/OL]. 2010-05-15. http://ictclas.org/Down_OpenSrc.asp

[9] 中国教育和科研计算机网紧急响应组. CCERT中文邮件数据集 CSDCE[DB/OL]. 2010-05-15. http://www.ccert.edu.cn/spam/sa/datasets.htm#4

[责任编辑: 顾晓天]