基于中文变形词匹配的贝叶斯邮件过滤模型

汪 霞 郑 宁 徐 明 陈 默
(杭州电子科技大学计算机学院 浙江 杭州 310018)

摘 要 针对特征词变异的中文垃圾邮件问题,提出了一种基于变形特征词匹配还原的新贝叶斯邮件过滤算法。改进的模型能自动发现邮件中的变异特征词,并根据对应的变异类型还原算法将其还原,避免了变异特征词的匹配逃脱。算法提高了对于含有拼音替换、同音字替换、符号插入等变形特征词样本的分类准确率。实验表明,改进的过滤算法比普通贝叶斯算法有更好的性能。

关键词 贝叶斯 垃圾邮件过滤 变形特征

BAYESIAN EMAIL FILTERING MODEL BASED ON CHINESE METAMORPHIC WORDS MATCHING

Wang Xia Zheng Ning Xu Ming Chen Mo
(School of Computer, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China)

Abstract This paper presents a new Bayesian email filtering algorithm based on metamorphic characteristic words matching and restoration against the problem of Chinese spam mail with characteristic words variation. The improved model can automatically detect varied characteristic words in the email, and restore them according to corresponding recovery algorithm for varied types which prevents the escape of the varied characteristic words from matching. The algorithm meliorates the classification accuracy of the samples of metamorphic characteristic words including Pinyin substitution, homophone substitution and symbols insertion, etc. The result of experiment shows that the improved algorithm has better performance than normal Bayesian algorithm.

Keywords Bayesian Spam mail filtering Metamorphic characteristic

0 引 言

随着互联网的迅速发展,电子邮件在方便人们交流的同时,随之而来的垃圾邮件也成为困扰人们的一大难题。如何准确地、自动地从大量邮件中过滤掉海量的垃圾邮件已成为一个重要的研究课题。

贝叶斯算法在垃圾邮件过滤技术中得到了很好的应用。它可以自动地对垃圾邮件进行学习,自动地适应用户的使用习惯,自动地过滤掉垃圾邮件,且准确率高。但是其动态适应性不强,对于包含如“免*费”、“mian费”这类变形特征词的变异邮件体,判别性不高,分类准确率变低。

针对这种情况,其他研究者也进行了研究,文献[1—3]提出基于人工免疫的垃圾邮件过滤系统,它将生物免疫原理引入贝叶斯分类器,使系统对一些特征词的变异体能够免疫,具有一定的自适应能力。文献[4]提出一种关键字匹配过滤方法,它采用基于语义单元表示树剪枝来过滤变形关键字,但要求事先收集一个庞大的关键字变形库。而文献[5]也提出了一种关键字匹配算法,同样要求事先收集变形关键词表。本文结合了贝叶斯和关键词匹配两种过滤技术的优点,提出了一种基于变形特征词匹配的贝叶斯邮件过滤算法。通过对原始贝叶斯过滤模型的改进,从训练期结束后提到的概率表中,自动生成变形特征词表;并针对拼音替换、同音字替换、符号插入等不同变形特征词的样本提出了变形词匹配算法;结合变形特征词表与变形词匹配算法进行变形词匹配,将变形特征词还原正常,再经贝叶斯分类器过滤,提高了变异垃圾邮件的过滤率。并用 CCERT 提供的中文邮件语料集做了对比实验与分析,取得了较好的效果。

1 改进的贝叶斯过滤模型

贝叶斯定理应用于邮件分类即是计算邮件 e_x 属于某个类别 C_j 的概率 $P(C_j|e_x)$,将邮件分到概率最大的类别中去。这里 C_j 分为合法邮件 C_l、垃圾邮件 C_s 两大类,计算 $P(C_j|e_x)$ 时,利用了贝叶斯公式:

$$P(C_j|e_x)=\frac{P(C_j)P(e_x|C_j)}{\sum_{j\in\{l,s\}}P(C_j)P(e_x|C_j)} \tag{1}$$

设 e_x 表示为特征集合 $(t_1, t_2, \cdots, t_n)$,n 为特征个数,且特征之间相互独立,则有:

$$P(e_x|C_j)=\prod_{i=1}^{n}P(t_i|C_j) \tag{2}$$

公式中需要先统计邮件集中单个特征词在某类邮件中出现的概率 $P(t_i|C_j)$,再求出邮件的特征联合概率 $P(e_x|C_j)$。本文采用 Paul Graham 提议的贝叶斯概率模型[6,7],主要维护三张哈希表:hash_bad、hash_good、spam_probability(简称 spam_p)。前

收稿日期:2008—06—03。浙江省自然科学基金项目(Y106176);浙江省科技厅计划项目(2007C33058)。汪霞,硕士生,主研领域:信息安全与信息处理。

两张分别存放垃圾、合法邮件中提取的特征及特征出现的次数，spam_p中存放邮件集中提取的特征和特征在垃圾邮件中出现的概率 $P(t_i|C_s)$。过滤时，每封邮件根据 spam_p计算邮件的特征联合概率进行分类。基于原始模型，本文提出了改进，在构造分类器之后，构造变形特征词表 hash_key，用以发现变异邮件中的变异特征词，再根据还原算法将词还原，重构邮件的文本向量，以提高贝叶斯过滤准确性。如图 1所示。

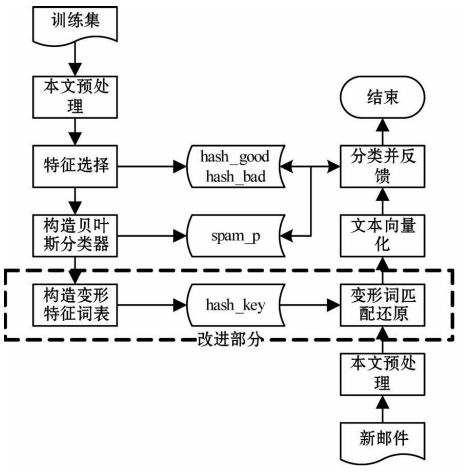

图 1 改进的邮件分类模型

2 构造变形特征词表

垃圾邮件之所以对某些特征进行变形，是因为这些词在 spam_p表中具有很高的概率值，而对于那些概率值低的特征，没必要进行变形，因此构造变形特征词表时，只要考虑概率值高的特征。将 spam_p表中概率值高于某个阈值（0.9）的特征取出，重新构造成一张变形特征词表 hash_key。hash_key中保存了特征 t_i与 t_i首字（或全词）的拼音。将 t_i首字（或全词）的拼音作为键值以便查寻，但可能有多个相同的键值，故采用多重映照哈希表结构进行保存。图 2为 hash_key的结构。

a	…	main	main	mianfei	…	…	z
…	…	免费	面试	免费	…	…	…

图 2 hash_key存储结构

3 中文变形特征词匹配

3.1 特征词的变形模式

垃圾邮件为了逃避过滤，将那些垃圾概率高的特征采用非正常形式出现，将一个词分解成若干汉字，以躲避特征匹配。特征词的变形模式主要为：

1）符号插入，如：优 &* 惠、法轮 … 功；

2）繁简体混合，如：美國；

3）拼音替换，如：免 fei，mianfei

4）同音字替换，如：尤惠；

5）图片替换。

3.2 文本预处理

提取出主题、正文、图片，计算图片的 MD5值。将主题与正文中的繁体转化为简体；统一英文大小写；保留中英文常用标点符号、数字，去除其余的非汉字字符，如“*”等。

分词后按文本内容顺序，将得到的特征依次存入数组 Array。如：“本月优*惠只要 200元，免 fei给用户更新”分词后得到的数组 Array如图 3所示。

本月	优惠	只要	200	元	,	免	fei	给	用户	…

图 3 Array数组

“*”在预处理时已被除去，不影响分词，故在数组 Array中“优惠”能够正常分词。由于垃圾邮件一般都是将垃圾概率高的特征拆分成字与拼音组合，匹配时只需对单个字或英文字符串进行变形特征匹配。

3.3 变形特征词匹配

对 Array进行匹配时，依次取其单元元素，遇到汉字词 t_i时，直接去 spam_p中匹配特征，获取特征 t_i对应的概率值 $P(t_i)$；遇到数字、空格、标点时，直接跳过；遇到单个汉字或英文字符串时，则从 Array当前位置依次提取单元元素，分别将元素与其对应的拼音存入临时字符串 Tstr与 Pstr，直至取到的元素是汉字词为止。如图 3所示，当取到元素“元”时，生成字符串 Tstr=“元，免 fei给”与 Pstr=“yuan，mianfeigei”。再根据之前构造的 hash_key，对 Tstr和 Pstr进行变形特征词匹配，若匹配成功，则根据还原的特征去 spam_p查询概率值，否则重新将当前单个汉字或英文字符串当作特征去 spam_p查询概率。处理完 Array中所有元素，计算出特征联合概率，即可对邮件分类。判断完毕后，修改 spam_p中的概率值，完成自学习。图 4为 Array匹配过程图。

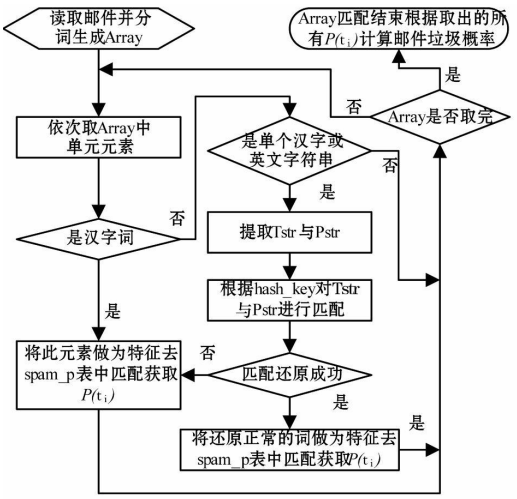

图 4 Array匹配流程图

此过程中，对 Tstr和 Pstr进行变形特征词匹配是关键，需要获取 hash_key表中相应特征 t_i的每个汉字符及汉字符对应的拼音，结合二者对 Tstr和 Pstr进行匹配。每个 t_i由若干个字符构成，假定 S_{ij}表示其中的一个字符，j代表第 j个字符。U_{ij}表示 S_{ij}对应的拼音，如：特征 t_i“免费”，它由两个字符构成，S_{i1}代表“免”，S_{i2}代表“费”，U_{i1}代表“mianr”，U_{i2}代表“fei”。

首先，需要判断遇到的元素是单个汉字还是英文字符串。若是单个汉字，则要先转成拼音，去变形特征词表 hash_key中查询，若匹配到，则获取此拼音对应的特征 t_i，并且计算 t_i包含的汉字个数 n，然后依次将此词对应的 n个 S_{ij}去 Tstr中查询匹配，若某一 S_{ij}匹配不到，则将此 S_{ij}对应的 U_{ij}去 Tstr中查询匹配。如果此词 n个字符的 S_{ij}或 U_{ij}都可以在 Tstr中匹配到，并且匹配字

符之间无汉字间隔,则认为匹配成功,将变形特征词还原为 t_i,flag置为 1,并标记最后一个匹配字符(S_{in}或 U_{in})的位置 pos,以便匹配还原后找到 Array中下一元素。若并非 n个字符的 S_{ij}或 U_{ij}都匹配中,但有超过一个字符的 S_{ij}或 U_{ij}匹配时,可能是同音字替换变形,需要再深入匹配。此次是将 t_i的所有字符 U_{ij}去 Pstr中查询匹配,若是全都匹配到,则认为匹配成功,还原后将 flag置为 1,同样标记位置 pos。这里要求至少有一个 S_{ij}或 U_{ij}匹配才进行深入匹配,是为了减少误判。具体算法如下:设初始 flag= 0,s= 0,u= 0(s与 u分别表示 S_{ij}与 U_{ij}匹配中的个数)。

算法 1　变形特征词匹配算法

```
If是单个汉字　Then
{ 字转化成拼音,去 hash_key中查找:
  If　匹配中　Then
  {　获取对应 ti与 tj的汉字个数 n;
    For (j= 0; j<= n; j++)
    {　If 在 Tstr中匹配到 Sij且与上个匹配
        字符之间无汉字间隔　Then
      {　s++; Continue }
      获取 Sij对应的拼音 Uij
      If　在 Tstr中匹配到 Uij且与上个匹配
        字符之间无汉字间隔　Then
      {　u++; Continue }
    }
    If (s+ u) == n　Then
      匹配成功, flag= 1,标记位置 pos;
    If s> 1或 u> 1　Then
    {　For (j= 0; j<= n; j++)
      If在 Pstr 中匹配到 Uij　Then
        Continue;
      If　j> n　Then
      匹配成功, flag= 1标记位置 pos;
    }
  }
}
```

遇到元素是英文字符串时,省去拼音转换一步,其余步骤与遇到单个汉字处理一致。汉字中的同音字很多,所以 hash_key中的键值可能会重复,匹配时需要设置一个循环,对表中同键值的项按上述方法一一匹配,flag为 1时,跳出循环。对于图片替换法,将图片的 MD5值作为特征,统计其垃圾概率。此方法对于用固定图片代替的特征,效果很好,若图片经常变换,效果不理想。

4　实验与评估

该实验的目的是为了比较改进后的模型与简单贝叶斯模型在邮件过滤准确性、自学习后的过滤时效性上的差异,因此分别实现这两个模型,方便进行结果对比。两个模型,都采用中国教育和科研网紧急响应组(CCERT)垃圾邮件数据库的公开语料库、同样的训练集和测试集。

4.1　实验环境搭建

实验自 CCERT提供的 2005年 6月数据库中随机挑选出 800封邮件作为训练集,其中垃圾邮件 400封,合法邮件 400封。CCERT提供的邮件有很多重复,考虑到实际情况,确实会在一段时期内收到很多相同的垃圾邮件,故对挑选中的重复邮件不作剔除,予以计算。测试集中的合法邮件都选自于 CCERT数据集,数量为 50封;测试集中的垃圾邮件用两种方式获取,一是从 CCERT数据集中选出垃圾邮件 50封,通过编程将这 50封邮件中垃圾概率大(大于 0.9)的特征词进行变形,根据不同的变形方式,得到两个变形垃圾邮件测试集 T1(多为字符插入)与 T2(多为拼音、同音字替换)。二是从自己邮箱中收集到的含有变形特征词的垃圾邮件中挑选 50封作为垃圾测试集 T3。

实验用 C++在 Linux平台上实现这两个模型,分词工具 ICTCLAS采用 Linux平台下的 1.0版本。分为四个阶段,首阶段从训练集中分别取垃圾、合法邮件各 100封作训练,并对测试集 T1、T2、T3进行测试与学习。以后每阶段在上一阶段训练学习的基础上,再从训练集中各取 100封邮件作训练,并完成三个测试集的分类与自学习。实验结果引入三个参数 R(Recall)、P(Precision)和 Acc(Accuracy)作为评估标准,如图 5—图 7所示。

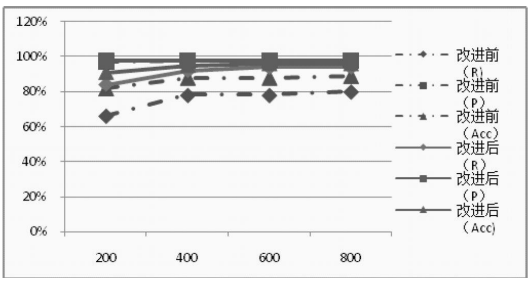

图 5　测试集 T1的实验对比结果

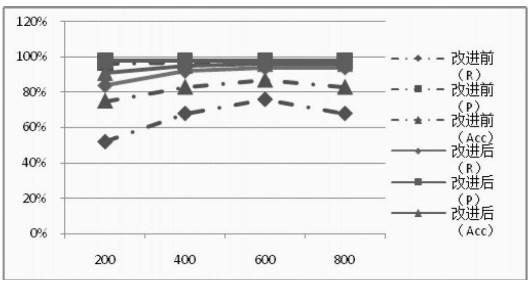

图 6　测试集 T2的实验对比结果

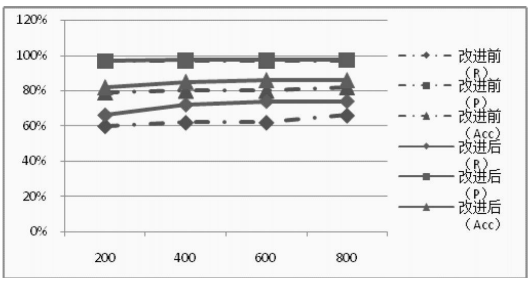

图 7　测试集 T3的实验对比结果

4.2　实验结果分析

从图 5—图 7中可以看出,改进过的方法优越性很明显。首先,改进过的模型比原简单模型在 R与 Acc值上有 3%—24%的提高,P值变化不明显,主要是改进后的方法只提高了垃圾邮件的检测率,对合法邮件的检测并无改进。其次,从过滤的时效性来看,由于原简单模型无法正确匹配变形特征词,需将其分解成新字词重新慢慢学习,故学习阶段变长。而改进过的模型,学习时间短,前几个阶段检测率就得到明显提高。第三,测试集 T3因为与训练集并非来自同个语料库,过滤效果不好,但改进后的效果比原有的好。最后,因为 T1经分词后得到的多是单个字,而 T2得到的多是拼音与同音字,由于在 spam_p中很少

(下转第 130页)

表 2 高维大定义域函数各算法最优结果对比

	SGA	FCGA	MSEP	改进的 GA	ACGA
F_1	0.147e-4	6.76e-4	2.2e-12	0.7424e-3	0
F_2	52.7328	19.3210	1.6e-8	5.1681e-2	0
F_3	6.5803	0.26e-2	4.0e-7	---	0
F_4	-9897	-11842	-11898.9	-12569.5	-12596
F_5	43.7345	57.0362	24.9	47.9772	0

表 3 多峰函数各算法结果对比

	SGA		FCGA		BFCGA		ACGA	
	Mean	best	Mean	best	Mean	best	Mean	best
F_6	34.34	3.35e-3	1.24e-6	5.21e-9	1.7e-1	0	1.45e-1	0
F_7	0.246	9.72e-3	0.0097	9.71e-3	9.5-2	0	1.01e-2	0

从表 1和表 3中可以看出算法 ACGA和 BFCGA都能够快速地找到函数的最优解，甚至可以找到精确解，这说明族间交叉算子和精英保留策略的变异算子使算法在进化过程中能快速收敛到最优解；其中 ACGA在平均适应度值指标上较 BFCGA更优一些；同时 ACGA在运行 100代即可找到最优解，NCE-GA虽然对 f_1也找到了最优解，但其运行代数为 2000，收敛速度低于本文算法；在表 2中各算法运行的代数与 MSEP[7]中的运行代数一致以便于结果的对比，但事实上 ACGA算法在运行 100代时同样可以找到最优解，这说明算法的收敛速度是相当快的，但此时它的平均适应度值较大；对于函数 f_4的优化结果可以看出，优化结果 -12596极其接近其近似值 -12596.5，说明 ACGA对于最小值不是 0的函数优化同样有效，也说明 ACGA在其它最小值为 0的函数优化中找到最优解 0不是偶然的。

4 结 论

本文创新点是将生长树聚类的思想应用于遗传算法中，并采用了新的遗传算子。通过对种群采用基于最小生成树的聚类方法聚类生成若干族，再采用相应的族间交叉算子和族内交叉算子及采用精英保留策略后的变异算子，即利用了聚类保持了种群多样性，又能达到快速收敛的效果。实验已证明了算法的有效性。在实验过程中我们发现：在解空间是一维的情况下，基于最小生成树的聚类方式与基于适应度的聚类方式结果是相同的；在理论上，ACGA算法对于有聚类特征的问题求解也应有较好的精度。这些还需实验进一步改进和证明。

参 考 文 献

[1] 徐立鸿，沈于晴. 一种基于家庭聚类思想的遗传算法[J]. 信息与控制，2004，33(5)：527-530.

[2] Martin Pelikan，David E Goldberg. Genetic algorithms，clustering，and the breaking of symmetry[C]. University of Illinois，Tech Rep 2000013，2000.

[3] 厍向阳，薛惠锋，高新波. 基于生长树的遗传聚类算法研究[J]. 计算机应用研究，2006(7)：62-64.

[4] Mana M A，Walcott B L. Stability and Optimality in Genetic Algorithm Controllers[C]//Dearborn：Proceedings of the 1996 IEEE International Symposium on Intelligent Control，MI-September 15-18，1996：492-496.

[5] Petra Kudova. Clustering Genetic Algorithm[C]//18th International Workshop on Database and Expert Systems Applications，1529-4188/07 DOI 10.1109/DEXA.2007.65：Computer Society：138-142.

[6] Norikazu IKOMA，Hiroshi MAEDA. Adaptive Order Selection with Aid of Genetic Algorithm[C]//Seoul，Korea：1999 IEEE International Fuzzy Systems Conference Proceedings August 22-25，1999：0-7803-5406-0/99 1999 IEEE：(III)：1785-1789.

[7] Hongbin Dong，Jun He，Houkuan Huang，Wei Hou. Evolutionary programming using a mixed mutation strategy[J]. Information Sciences，2007，177：312-327.

[8] Swagatam Das，Ajith Abraham，Amit Konar. Automatic Clustering Using an Improved Differential Evolution Algorithm[C]//IEEE Transactions on Systems，Man，and Cybernetics-Part A：Systems And Humans，2008，38(1)：218-237.

[9] 郑金华，史忠植，谢勇. 基于聚类的快速多目标遗传算法[J]. 计算机研究与发展，2004，41(7)：1081-1087.

[10] 陆林花，王波. 一种改进的遗传聚类算法[J]. 计算机工程与应用，2007，43(21)：170-172.

[11] 陈有青，徐蔡星，钟文亮，等. 一种改进选择算子的遗传算法[J]. 计算机工程与应用，2008，44(2)：44-50.

[12] 姚金涛，杨波. 一种具有自然血亲排斥的遗传算法研究[J]. 计算机工程与应用，2008，44(16)：27-29.

(上接第 107页)

出现拼音，T2比 T1难以匹配，故原有模型下 T1比 T2实验结果要好，但在改进模型下两者都被还原正常，故实验结果一致。

5 结 论

本文提出了一种新垃圾邮件过滤算法。这种算法有多个优点：它提高了对变形词的识别能力，缩短了训练学习时间；同时在匹配过程中，灵活地根据原有的贝叶斯模型，自动生成与邮件相关性强的变形特征词表帮助匹配，解决了人工收集特征词表的相关性差、费时多等问题。实验结果证明，检测效率得到了提高，学习速度变快，这对于开发针对变异邮件的过滤模型具有较强的启发意义。下一步研究中将针对其它变异模式的变异特征(如图片替换、字拆分替换等)进行识别与还原，进一步提高过滤效率。

参 考 文 献

[1] 秦志光，罗琴，张凤荔. 一种混合的垃圾邮件过滤算法研究[J]. 电子科技大学学报，2007，36(3)：385-388.

[2] 张成功，黄迪明，胡德昆. 基于人工免疫原理的反垃圾邮件系统AIASS[J]. 电子科技大学学报，2007，36(1)：96-99.

[3] 匡胤，黄迪明. 基于抗体网络的邮件过滤器设计[J]. 电子科技大学学报，2006，35(5)：810-813.

[4] 高庆狮，李莉，刘宏岚. 基于语义单元表示树剪枝的关键字过滤方法[J]. 北京科技大学学报，2006，28(12)：1191-1195.

[5] 范黎林，王晓东. 一种用于垃圾邮件过滤的中文关键词匹配算法[J]. 河南科技大学学报，2006，27(5)：35-37.

[6] Paul Graham. A plan for Spam[EB/OL]. 2002-08. http://paulgraham.com/spam.html.

[7] Paul Graham. Better Bayesian Filtering[EB/OL]. 2003-01. http://www.paulgraham.com/better.html. January 2003.