一种基于成词概率的贝叶斯垃圾邮件过滤方法

林 伟
(四川警察学院 计算机系,四川 泸州 646000)

摘 要:贝叶斯分类方法在英文邮件过滤中效果良好,在中文环境下一直表现不佳,而特征选择是垃圾邮件过滤中的重要步骤,它能够有效地改善过滤效果。文中以成词概率作为特征选择的基础,用构造的方法形成候选特征集,然后进一步用信息增益的方法来度量特征与类的关系,选择信息增益较大的N个特征做为最后的特征向量空间。在此基础上利用贝叶斯方法对邮件进行分类,实验结果验证了该方法在分类时间和分类效果上都优于传统的基于机械分词的贝叶斯方法。

关键词:垃圾邮件;成词概率;贝叶斯方法
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2011)09-0242-03

A Bayesian Spam Filtering Method Based on Words Probability

LIN Wei
(Department of Computer Science, Sichuan Police College, Luzhou 646000, China)

Abstract: Bayesian classification method has expressed high accuracy in English mails filtration, but the performance was not good under Chinese environment. It has taken the words probability as the foundation of the feature selection, the candidate feature sets were formed through the construction method, then use information gain to evaluate the relationship between feature and class, choose the n-larger information gain features as the final feature vector space. Based on this, the mails were classified by Bayesian method. Experimental verification shows this method surpassed the tradition method which based on the mechanical participle of the Bayesian theorem in the classified time and the classified effect.

Key words: spam; words probability; Bayesian method

0 引 言

随着互联网的快速发展,电子邮件以其迅捷、方便、低成本的优点而得到广泛应用。与此同时,日益泛滥的垃圾邮件严重影响了电子邮件用户的正常使用,给网络安全带来了极大的威胁。中国互联网协会2008年度第一次反垃圾邮件调查报告指出[1],中国网民平均每周收到17.64封垃圾邮件,占所有邮件的56.70%。每个中国网民平均每周用于处理垃圾邮件所消耗的时间为12.11分钟,以单位时间GDP折算,垃圾邮件每年带来的损失超过40亿元人民币。反垃圾邮件技术已经成为了相关领域内的研究热点问题。

在目前的反垃圾邮件技术研究中,基于内容的过滤技术依然是研究的重点和主流。基于内容的邮件过滤本质上是一个二元分类问题,即将到达的邮件分为垃圾邮件和合法邮件两大类,去掉标记为垃圾邮件的一类。常用的分类方法包括贝叶斯方法、SVM、KNN、Boosting等,其中贝叶斯方法由于其分类效果好、分类时间短而得到广泛应用。Sahami[2]最早将简单贝叶斯方法用于邮件进行过滤,在Sahami的实验中90%以上的英文垃圾邮件能够被过滤掉。Androutsopoulos[3]和Graham[4]同样用简单贝叶斯方法过滤垃圾邮件,正确率能够达到99.5%,而误判率几乎为0。

虽然贝叶斯方法在英文邮件过滤中较为成功,但在中文邮件过滤中的表现则不太理想。主要原因是由于中英文构词和语法上存在巨大差异。中文词语之间没有分隔符,难以确定语言单元(Token)的长度。中文语法结构较为松散,词语之间的关系复杂,这些都为贝叶斯方法应用于中文邮件过滤增加了难度。

针对这个问题,提出了一种基于成词概率的贝叶斯邮件过滤方法。该方法根据训练集合中Token单元的词频来计算词语的成词概率,通过成词阈值参数来控制成词准确度以及计算复杂度。在此基础上利用贝叶斯分类模型实现垃圾邮件过滤。实验结果表明,该

收稿日期:2011-04-07;**修回日期**:2011-07-10
基金项目:四川省青年软件创新工程基金(2007AA42)
作者简介:林 伟(1983-),男,讲师,研究方向为数据挖掘与机器学习、网络安全。

方法能够有效提高传统贝叶斯过滤方法的过滤精度与处理速度。

1　基于成词概率的特征选择

1.1　邮件表示

邮件属于半结构化的文本,计算机不能直接处理,需要转换成可计算的形式。通常采用信息处理中广泛使用的向量空间模型[2](Vector Space Model,VSM)对邮件进行向量化表示。

在向量空间模型中,一封邮件可以表示成 $d_i(V,i,C)$ 的形式,其中 V 为邮件的特征向量,i 为邮件编号,C 为类别标志,1 表示合法邮件,0 表示垃圾邮件。记邮件的特征空间为 $\Omega(w_1,w_2,\cdots,w_n)$,$w_i$ 为邮件的一个特征,邮件 d 的特征向量表示为 $V(x_1,x_2,\cdots,x_n)$,x_i 为邮件 d 在特征向量空间中对应特征的测度 。已知训练集 $D=\{d_1,d_2,\cdots,d_N\}$ 中所有邮件的类别,邮件过滤的任务就是将新到达的邮件进行类别标记。邮件过滤问题本质上就是一个有监督的分类问题。

1.2　特征选择

特征选择(Feature Selection)指的是从原有的 M 个特征集合中选择 N 个可分性较好的特征子集,从而降低维度,提高分类效果。特征选择是模式识别和数据挖掘领域的重要问题之一,同时也是难以处理的一个问题。采用穷举遍历特征空间的所有特征的可能组合,从而选取最优的特征子集的方法能在一定程度上得到最优子集[5],但在实际应用时由于特征空间较大,计算时间代价和复杂度太大,因此可行性不强。为了折中分类的性能和计算代价,给出了一种成词概率的特征子集的构造方法,首先从原有的特征集合中选择成词概率较大的特征组成新特征空间,然后从中选择 N 个特征构成特征子集。

1.2.1　候选特征集

根据向量空间模型,对于任意一封邮件,可以形式化描述为 $E=<w_1,w_2,w_3,\cdots,w_n>$,则原始特征空间定义为参与训练的所有邮件的 W 组成的集合。

令 T_i 为特征全集中长度为 i 的全部特征集合,描述为 $T=\{T_1,\cdots,T_i,\cdots\}$。因为对于有监督的分类问题而言,低频特征属于噪声数据,会影响分类效果[6]。因此,首先去掉低频特征,构成新的特征集合 $T'=\{T_1',\cdots,T_i',\cdots\}$。下面给出利用成词概率构造新特征的过程。

定义　令长度为 i 的非低频特征集合 $T_i'=\{w_i[1],\cdots,w_i[j],\cdots,w_i[|W_i|]\}$,$w_i[j]$ 表示 W_i 中的第 j 个特征。已知非空集合 W_1,记 $w_i[j]$ 与 $w_1[k]$ 组成的新特征为 w_{i+1}。$C(w_{i+1}\mid w_i[j])=\mathrm{TF}(w_{i+1})/\mathrm{TF}(w_i[j])$ 表示 w_{i+1} 的成词概率。

给定成词概率阈值 $0<\alpha<\beta<1$,如果

(1) $C(w_{i+1}\mid w_i[j])\leqslant\alpha$,则表示 w_{i+1} 是特征词的概率较小,w_{i+1} 不放入 W_{i+1} 中,取 $w_1[k+1]$ 与 $w_i[j]$ 组合。

(2) $\alpha<C(w_{i+1,k+j}\mid w_{i,k})<\beta$,则表示 w_{i+1} 是候选特征,将 w_{i+1} 放入 W_{i+1} 中,取 $w_1[k+1]$ 与 $w_i[j]$ 组合。

(3) $C(w_{i+1}\mid w_i[j])\geqslant\beta$,则表示 $w_{1,j}$ 大多数情况下都紧随 w_{i+1} 出现,w_{i+1} 可以替代 $w_{i,k}$。删去 W_{i+1} 中由 $w_i[j]$ 生成的其他特征。将 w_{i+1} 放入 W_{i+1} 中,取 $w_i[j+1]$ 与 $w_1[1]$ 组合。

根据以上步骤逐步构造 $W_1,W_2,\cdots,W_i$。显然,随着 i 的增加,W_i 的规模会迅速减小,直至为空。候选特征集 $\Omega_c=\{W_1,W_2,\cdots,W_i\}$。

1.2.2　特征筛选

在文本分类领域,为了量化特征 w_i 与类别之间的共享信息,常用的方法有信息增益(IG)、文档频率(DF)、互信息(MI)、CHI 检验(CHI)等。Yang 和 Pedersen 等对目前常用于文本分类的几种特征选择方法做出了比较全面的研究,指出 IG 和 CHI 比 MI 和 DF 具有更好的效果[7]。因此对于初选的特征集 Ω_c 中的所有特征,采用信息增益进行特征选择。

信息增益在机器学习领域应用较为广泛,在信息论中,样本属性的信息增益越大,其所包含的信息量也将越大。给定一个词语的信息增益定义如下:

$$IG(w_i)=-\sum_{j=1}^{k}P(c_j)\log P(c_j)+P(w_i)\sum_{j=1}^{k}P(c_j\mid w_i)\log P(c_j\mid w_i)+P(\bar{w_i})\sum_{j=1}^{k}P(c_j\mid\bar{w_i})\log P(c_j\mid\bar{w_i}) \tag{1}$$

其中,$P(c_j)$ 为第 j 类别出现的概率;$P(c_j\mid w_i)$ 为包含特征词 w_i,并且属于第 j 类文档的概率;$P(c_j\mid\bar{w_i})$ 为不包含特征词 w_i,并且属于第 j 类文档的概率。

对于所有初选的特征,分别计算其信息增益,选择信息增益值较大的 n 个特征构成最后的特征向量空间 $\Omega(w_1,w_2,\cdots,w_n)$[8,9]。

2　贝叶斯分类模型

朴素贝叶斯分类模型是利用类别的先验概率和词对于类别的条件概率来计算未知文本属于某一类别的概率,它是建立在"贝叶斯假设"基础上,即假定所有特征之间相互独立。应用到邮件分类中,它是通过一定数量的垃圾邮件(Spam)和非垃圾邮件(Ham)做为邮件训练集来训练出分类模型,将训练的结果作为判定未知邮件类别的主要依据。

给定一封未知类别的邮件,使用向量空间模型(VSM)对其形式化描述为 $E=<w_1,w_2,w_3,\cdots,w_n>$,

其中特征属性 $< w_1, w_2, w_3, \cdots, w_n >$ 之间假设相互独立。邮件分类器的任务就是计算出待分类邮件是垃圾邮件的概率，如果它超过某一个阈值则认为该邮件为垃圾邮件。对于该邮件 E 根据贝叶斯理论，相应类别的概率按公式(2)计算：

$$P(c \mid E) = \frac{P(c) \cdot P(E \mid c)}{\sum_{c \in C} P(c) \cdot P(E \mid c)} \tag{2}$$

其中 $c \in C = \{\text{Spam}, \text{Ham}\}$ 该公式中 $P(c \mid E)$ 表示邮件 E 属于类别 c 的概率。$P(E \mid c)$ 表示假定类别为 c 时邮件的先验概率，$P(c)$ 表示类别 c 的先验概率，$P(e)$ 对于同一个邮件不变。若 $P(\text{Spam} \mid E) > P(\text{Ham} \mid E)$ 时，即认为该邮件 E 为垃圾邮件，否则为正常邮件[10,11]。

3 分类实验

3.1 实验数据

文中实验数据全部采用中文邮件。实验数据为 CCERT 提供的中文邮件数据集及笔者收集的部分私人邮件。实验数据的样本结构如表 1 所示。从垃圾邮件和合法邮件中各取 200 封作为测试样本，其余的作为训练样本。

表 1 样本结构

来源	合法邮件	垃圾邮件
CDSCE	1400	1500
自采集样本	300	200
合计	1700	1700

3.2 评价标准

通常采用召回率(recall)、误判率(FP)和 TCR[12] 评价邮件过滤系统的过滤性能。设测试集合中共有 N 封邮件，其中垃圾邮件为 N_s 封，合法邮件 N_h 封。记 $N_{s,s}$ 表示系统将垃圾邮件正确判断为垃圾邮件的数量，$N_{s,h}$ 表示将垃圾邮件判断为合法邮件的数量，$N_{h,s}$ 表示将合法邮件判断为垃圾邮件的数量。

$$\text{Recall} = \frac{N_{s,s}}{N_s} \times 100\%$$

$$\text{FP} = \frac{N_{h,s}}{N_h} \times 100\%$$

$$\text{TCR} = \frac{N_s}{N_{h,s} + N_{s,h}} \times 100\%$$

Recall 值反映了过滤系统对垃圾邮件的敏感程度，通常情况下，提高 Recall 值的同时 FP 值也会随之增大。TCR 则反映了过滤系统对邮件系统的改进能力。在实际应用中，过滤系统的处理速度同样十分重要。用训练时间 T 和单封邮件分类速度 t 来衡量系统的处理速度。

3.3 实验环境

实验环境：CPU P4 2.4G，内存 2G，硬盘 80G，操作系统为 Windows Server 2000。开发环境使用 Microsoft VC++6.0 平台。

4 实验分析

4.1 实验一

训练样本共 3000 封，其中合法邮件 1500 封，垃圾邮件 1500 封。只保留每封邮件的标题和正文字符串，邮件的平均长度为 187.6(字符)。表 2 为提取 1000 个互信息最大的目标特征的实验结果。分别列出了在参数 λ、α、β 的不同取值下，候选特征词数 $\sum |W_i|$、平均特征长度 $\overline{|w|}$ 和训练时间 T 的实验结果。

表 2 特征提取结果

(λ, α, β)	$\sum \lvert W_i \rvert$	$\overline{\lvert W \rvert}$	T(s)
(1, 0, 0)	87367	2.461	1106.4
(3, 0.1, 0.9)	17917	2.232	190.1
(5, 0.1, 0.9)	10611	2.121	171.0
(5, 0.2, 0.8)	9932	2.094	150.0
(7, 0.2, 0.8)	5429	1.746	125.5

当 $\lambda = 1, \alpha = 0, \beta = 0$ 的时候(即不考虑成词概率)，候选词条特征数接近 10 万条。从 87367 个候选特征中选出 1000 个目标特征的训练时间是比较长的。

从表 2 可以看出，随着 λ 的增大和 (α, β) 区间的缩小，候选词条数会迅速减小。当 $(\lambda = 7, \alpha = 0.2, \beta = 0.8)$ 时，候选特征只有 5429 条。λ 过大或者 (α, β) 区间过小都会使得部分有用的特征被忽略掉。通过后面的分类实验发现，在分类器参数相同的情况下，成词参数 $(\lambda = 5, \alpha = 0.2, \beta = 0.2)$ 的分类效果要优于 $(\lambda = 1, \alpha = 0, \beta = 0)$ 的分类效果。

4.2 实验二

取 $\eta = 1.2$，测试参数 λ、α、β 的分类结果指标如表 3 所示。

表 3 分类效果(η=1.2)

参数	指标			
$(\lambda, \alpha, \beta, n)$	n	Recall	FT	TCR
(1, 0, 0, 1500)	1500	63.2%	13.1%	1.32
(5, 0.1, 0.9, 800)	800	93.4%	0.56%	2.31
(5, 0.2, 0.8, 800)	800	94.2%	2.20%	2.54
(7, 0.2, 0.8, 800)	800	65.2%	17.2%	1.07

从表 3 可以看出，$(\lambda = 5, \alpha = 0.2, \beta = 0.8)$ 的分类效果最好，而且只需要 800 个特征就能达到 2.54 的 TCR 值，对系统的提升效果明显。特别需要说明的是参数(5, 0.1, 0.9, 800)与(5, 0.2, 0.8, 800)的分类效果反而是训练时间较短的(5, 0.2, 0.8, 800)效果更

(下转第 249 页)

队所需时间;

T_{com}为通信时间,可用如下公式估算:

$TC(X)=C0+C1*X$

其中:X为数据的传输量,通常以bit为单位;

C0为两站点间通信初始化一次所花费的时间,由通信系统确定,近似一个常数,以秒为单位;

C1为传输率,即单位数据传输的时间,单位是b/s。

4 结束语

简单的分布式封锁方法、主副本封锁方法和完全分布式加锁算法对数据对象的封锁所需通信开销大、锁管理复杂、事务并发难度大。通过分析分布式数据库加锁管理算法,利用全局目录和事务调度器,提出了一种新的分布式数据库加锁管理算法,为分布式数据库系统开发提供了一种新的思路。

参考文献:

[1] Ceri S, Navathe B, Wiederhold G. Distribution Design of Logical Database Schemas[J]. IEEE Transactions on Software Engineering, 1983, 9(1): 156-170.

[2] Jin Jing, Bukhres O, Elmagarmid A. Distributed Lock Management for Mobile Transactions[C]//15th International Conference on Distributed Computing Systems. Vancouver, BC, Canada: [s. n.], 1995: 243-251.

[3] Kedem Z M, Silberschatz A. Locking Protocols: From Exclusive to Shared Locks[J]. Journal of the ACM, 1983, 30(4): 15-24.

[4] Yannakakis M, Papadimitrioum C H, Kung H T. Locking Protocols: Safety and Freedom from Deadlock[C]//Proceedings of the IEEE Symposium on the Foundations of Computer Science. [s. l.]: [s. n.], 1979: 43-55.

[5] Wiesmann M, Pedone F, Schiper A, et al. Understanding replication in databases and distributed systems[C]//In Proc. of the 20th International Conference on Distributed Computing Systems (ICDCS). Taiwan: IEEE-CS Press, 2000: 464 - 474.

[6] Badal D Z. The distributed deadlock detection algorithm[J]. ACM Trans. Comp. Syst., 1986, 4(4): 320-337.

[7] Thomasian A. Two-phase locking performance and its thrashing behavior[J]. ACM Trans Database Syst, 1993, 18(4): 579-625.

[8] Gray J N, Lorie R A, Putzolu G R. Granularity of Locks and Degrees of Consistency in a Shared Data Base[J]. In Proceedings of VLDB, 1975, 3(8): 331-340.

[9] Ozsu M T. Principles of Distributed Database Systems (2E)[M]. 北京:清华大学出版社, 2002: 120-135.

[10] 宁　伟,李　艳,翟桂丹,等. 分布式数据库加锁与刷新机制的研究[J]. 内蒙古师大学报, 2001, 6(2): 123-126.

[11] 陈建英,刘心松,谈文蓉,等. 全局目录的动态管理和维护[J]. 计算机工程, 2006, 32(13): 35-37.

(上接第244页)

好。可能原因是成词区间减少了干扰特征的生成。

5 结束语

文中提出的基于成词概率的贝叶斯方法不仅能够减少训练时间和候选特征维数,而且在较小的特征空间中能够得到更优的分类效果。稍加修改还可以用于多类别的文本在线分类问题。

由于目前国内还没有权威的垃圾邮件语料库可供研究者使用,本方法只是在已有的数据集上有较好的效果。在实际应用中,不同的环境下参数设置会有诸多变化,能否经得起检验还有待进一步的证明。

参考文献:

[1] 中国互联网协会反垃圾邮件中心. 中国互联网协会2008年度第一次垃圾邮件调查报告[EB/OL]. 2008. http://www.anti-spam.cn/pdf/2008_1_dc.pdf.

[2] Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering junk e-mail[C]// AAAI-98 Workshop on Learning for Text Categorization. [s. l.]: [s. n.], 1998.

[3] Androutsopoulos I. An Evaluation of Naive Bayesian Anti-Spam Filtering[C]// Proc. of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning. [s. l.]: [s. n.], 2002.

[4] Graham P. A Plan for Spam[EB/OL]. 2002. http://www.paulgraham.com/spam.html.

[5] 崔自峰,徐宝文,张卫丰,等. 一种近似Markov Blanket最优特征选择算法[J]. 计算机学报, 2007, 30(12): 74-81.

[6] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004, 18(1): 26-32.

[7] Yang Yiming. A Comparative Study on Feature Selection in Text Categorization[C]//The ICML97. Nashville: [s. n.], 1997.

[8] 董　梅,胡学钢. 基于多特征选择的中文文本分类[J]. 计算机技术与发展, 2007, 17(7): 117-119.

[9] Han Jiawei. 数据挖掘概念与技术[M]. 北京:机械工业出版社, 2001.

[10] 詹　川. 反垃圾邮件技术研究[D]. 成都:电子科技大学, 2005.

[11] 陈晋川,陈治璋,贾洪明,等. 基于模式的贝叶斯垃圾邮件过滤的研究与实现[J]. 计算机工程与应用, 2006(6): 172-175.

[12] 王　申. 基于内容的垃圾邮件过滤技术若干研究[D]. 北京:中国科学院, 2005.