Naive Bayesian 算法在基于内容的垃圾邮件过滤中的应用①

Research on content - based anti - spam filtering using Naive Bayesian

李 欣 左瑞欣 曲文斌 (石家庄经济学院 050031)

摘要:垃圾邮件是 Internet 上面临急待解决的问题。Naive Bayesian 算法由于其简单高效性在文本分类中应用较广,重点阐述了 Naive Bayesian 算法在基于内容的垃圾邮件过滤中的应用,并在 Ling - spam 语料库上进行了实验,获得了良好的分类效果,实验论证了它在垃圾邮件过滤中的可行性和有效性。

关键词:垃圾邮件 Naive Bayesian 算法 查准率 查全率

1 引言

随着 Internet 的普及,电子邮件已成为人们日常交流的重要手段之一,但垃圾邮件问题也日益严重。据 2005 年 7 月中国互联网信息中心(CNNIC)发布的《第十六次中国互联网络发展状况报告》显示,用户平均每周收到 14.5 封电子邮件,其中垃圾邮件占 9.3 封。迄今为止,垃圾邮件没有统一的定义,《中国互联网协会反垃圾邮件规范》中指出,垃圾邮件为具有下述属性的电子邮件:

(1) 收件人事先没有提出要求或者同意接收的广告、电子刊物、各种形式的宣传品等宣传性的电子邮件;

(2) 收件人无法拒收的电子邮件;

(3) 隐藏发件人身份、地址、标题等信息的电子学邮件;

(4) 含有虚假的信息源、发件人、路由等信息的电子邮件[2]。Internet 上常见的名词 spam、UCE(Unsolicited Commercial Email 不请自来的商业电子邮件)UBE(Unsolicited Bulk Email 不请自来的批量电子邮件)与通常所说的垃圾邮件是一样的。这些垃圾邮件在占据邮件服务器大量存储空间的同时,也花费了用户大量的时间与金钱,降低网络运行效率,占用网络带宽,干扰邮件系统的正常运行,而且对网络安全造成极大的威胁。

面临垃圾邮件日益严重的现状,防范垃圾邮件的手段也应运而生,一般为:

(1) 反垃圾邮件立法。但立法面临一系列的问题,首先是垃圾邮件的概念之争,到底什么是垃圾邮件,像宣传品、电子期刊等是不是垃圾邮件? 其次是法律的执行问题,给予什么样的处罚? 由于缺少国际间的合作,即使发现来自境外的垃圾邮件,也无法制裁;如果规定发邮件需要额外的代价,目前很难被广大的邮件用户认可。

(2) 利用垃圾邮件过滤技术。近年来,有关垃圾邮件过滤的技术开始逐步兴起,如果能从技术上解决垃圾邮件问题,那将是最理想的。

2 垃圾邮件过滤技术

目前,防范垃圾邮件的主要手段为黑名单与白名单技术、基于规则的过滤技术、基于内容的过滤技术。

黑名单与白名单技术:黑名单中的发件人发送的邮件都认为是垃圾邮件,白名单中的发件人发送的邮

① 项目:河北省科技攻关计划项目,项目编号:05213568

件都认为是合法邮件，这是目前电子邮件过滤中广泛使用的技术。通常做法是收集一个黑、白名单列表，可以是电子邮件地址，也可以是邮件服务器的域名、IP 地址，当收到邮件时对发件人进行实时检查。这种名单一般由有信誉的组织提供，如中国互联网协会定期在主页上公开垃圾邮件服务器 IP 地址名单，个人也可根据需求定义和维护自己的黑、白名单。由于发送者在不断地变化，而且目前国内使用黑名单服务的邮件商很少，因此黑、白名单技术有局限性。

基于规则的过滤技术：如 outlook express 等客户端邮件系统，可按用户设定的过滤规则工作，但是，这种邮件过滤的方法以人为中心，且这种基于关键字比较的规则过滤技术具有以下缺点：要求用户有较强的专业知识、丰富的经验和充裕的时间去建立这些过滤规则，且过程很繁琐；if then 规则形式表达能力单一，只能做出是或不是的判断，推理缺乏灵活性；用于比较的关键字集合太小，不具备对邮件进行整体分析的能力；不适应动态变化，由于垃圾邮件具有发件人地址随机变化，邮件主题随机变化，伪造邮件头干扰信息等特点，用户必须经常调整这些规则，这给用户带来了较大的工作量。

基于内容的过滤技术：由于垃圾邮件注重个性化，判别的准则会随着时间变化，垃圾邮件本身的内容形式也在不断变化，通过对电子邮件的内容进行分析，识别出垃圾邮件，基于内容的过滤技术提供了更为准确的邮件过滤方法，这是本文的研究重点。

3 Naive Bayesian 算法在基于内容的垃圾邮件过滤中的应用

Naive Bayesian 分类算法提供了一种简单、高效的基于内容的分类方法。其原理是计算文本 d_x 属于某个类别的概率。

已知邮件的文本集 $D=\{d_1,d_2,\cdots d_n\}$ 和它的词汇集 $W=\{w_1,w_2,\cdots w_m\}$。采用 VSM 表示邮件，即：$\vec{V}(d_i)=(val(w_1),val(w_2),\cdots,val(w_m))$，$ci\in C=\{c_1,c_2,\cdots c_k\}$ 是类别变量。分类的任务就是对未知类别的邮件，来预测它的类别：$c=\max\{P(c_i|d_i)\}$，依据 Naive Bayesian 假设，邮件中各个特征项相对于类别属性是相对独立的，则 Naive Bayesian 公式为：

$$P(c_i|d_i)=\frac{P(c_i)\prod_{w_t\in d_i}P(w_t|c_i)}{P(d_i)} \quad (1)$$

对同一封邮件文本，$P(d_i)$ 不变。

$P(c_i)$：类的先验概率，由训练集估计。

$$P(c_i)=\frac{\text{训练集中属于 } c_i \text{ 类的文本数量}}{\text{训练集中的文本总数量}} \quad (2)$$

$P(w_t|c_i)$：特征 w_t 在类别 c_i 中出现的概率

在训练集上估计，为了简化计算的复杂性，采用最大似然估计（m－estimate）来估计 $P(w_t|c_i)$ 的值。

$$P(w_t|c_i)=\frac{1+n_i}{n+|W|} \quad (3)$$

n_i：c_i 类的所有文本中特征 w_t 的出现次数

$|W|$：c_i 类的所有文本中出现的特征总数

由于邮件类别只有垃圾邮件和合法邮件两类，直接使用 Naive Bayesian 算法分类偏差较大，再者，人们无法容忍合法邮件被误判为垃圾邮件而被误杀。为避免误判，当一封邮件被判为垃圾邮件时，需满足

$$\frac{P(c=spam|d)}{P(c=legitimate|d)}\geq\lambda \quad (4)$$

其中，$P(c=spam|d)=1-P(c=legitimate|d)$，公式（4）等价于

$$P(c=spam|d)\geq t \quad \left[t=\frac{\lambda}{1+\lambda}\right] \quad (5)$$

依据实验结果，当 $t=0.999$（$\lambda=999$）时，过滤结果是比较准确的。

Naive Bayesian 分类算法的训练流程如图 1 所示，分类流程如图 2 所示。

4 实验结果及性能评价

4.1 语料库

为了对 Naive Bayesian 分类算法在垃圾邮件过滤中的性能评估，需要一个公共的语料库，包含训练集和测试集。本系统采用的 Ling－Spam 英文语料库由希腊的 Androutsopoulos 等人提供，可以从 http://iit.demokritos.gr/skel/i－config/downloads/ 下载。Ling－Spam 语料库包含 481 封垃圾邮件和来自于语言学家列表的 2412 封合法邮件。Ling－Spam 的邮件来自公用邮件列表，因此，邮件内容没有加密。Ling－Spam 分为 10 份，每份大约 289 封邮件。Ling－Spam 语料包含 4 种形式。Bare 语料：没有用 Lemmatiser，也没有去除停用词；lemm 语料：使用了 Lemmatiser，但没有去除

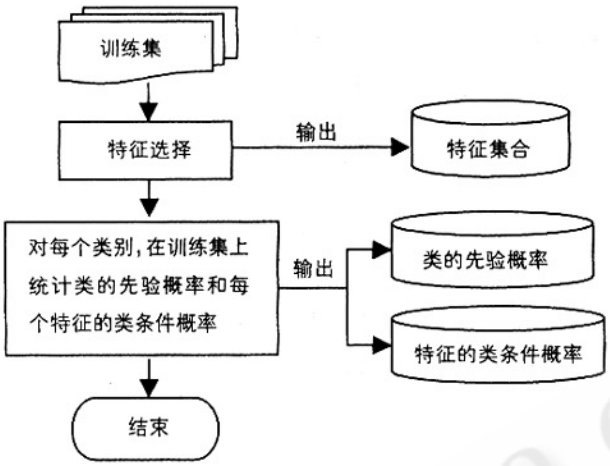

图 1　Naive Bayesian 分类算法的训练流程

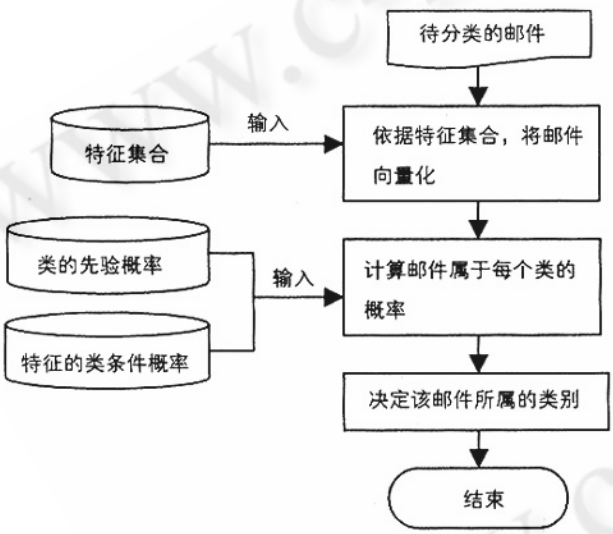

图 2　Naive Bayesian 分类算法的分类流程

停用词；stop 语料：没有用 Lemmatiser，去除了停用词；lemm_stop：使用了 Lemmatiser，也去除了停用词。实验采用预留法对 Naive Bayesian 分类算法在垃圾邮件过滤中的性能进行评估，即用训练集来构造基于内容的垃圾邮件过滤器，然后使用这个过滤器对测试集中的邮件进行分类。

4.2 评价指标及实验结果

常用的评价指标有分类正确率、查准率、查全率。通过表 1 的邻接表给出每个指标的定义。

表 1　邻接表

	垃圾邮件	合法邮件
系统判定为垃圾邮件	A	B
系统判定为合法邮件	C	D

分类正确率（Accuracy）：$Accuracy = \frac{A+D}{A+B+C+D} * 100\%$，对所有邮件的判对率。

查准率（Precision）：$precision = \frac{A}{A+B} * 100\%$，垃圾邮件的检对率，反映了过滤器找对垃圾邮件的能力。

查全率（Recall）：$Recall = \frac{A}{A+C} * 100\%$，垃圾邮件检出率，反映了过滤器发现垃圾邮件的能力[6]。

表 2 为使用预留法在 Ling－Spam 语料上的实验结果，可见将 Naive Bayesian 分类算法应用于垃圾邮件过滤中具有较高的分类正确率与查准率。

表 2　实验结果

使用的语料	分类正确率（100%）	查准率（100%）	查全率（100%）
Ling－Spam bare	87.2	93.2	72.6
Ling－Spam lemm	86.4	92.2	71.8
Ling－Spam stop	89.4	93.6	78.1
Ling－Spam lemm_stop	89.2	92.3	78.3

5　结束语

使用 Naive Bayesian 算法，依据邮件的内容进行智能过滤，从而将垃圾邮件与合法邮件分开，实验论证这种方法是高效、可行的。后续工作是进一步提高邮件过滤的准确率和查全率，并将该过滤器应用到实际的电子邮件系统中。

参考文献

1　第十五次中国互联网络发展状况调查统计报告［EB/OL］. http://www.cnnic.net.cn/html/Dir/2005/01/18/2744.htm. 2005.

2　中国互联网反垃圾邮件联盟［EB/OL］. http://www.anti－spam.org.cn.

（下转第 54 页）

（上接第 50 页）

3 Mehran Sahami, Susan Dumais, David Heckerman et al. A Bayesian Approach to Filtering Junk E-mail [R]. Learning for Text Categorization: Papers from AAAI Workshop. Madison, Wisconsin, 1998,55-62.

4 I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, C. D. Spyropoulos. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Encrypted Personal E-mail Messages [R]. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), Athens, Greece, 20001,60-167.

5 史忠植，知识发现[M]，北京 清华大学出版社，2002。

6 王斌，文本分类综述[EB/OL]，http://159.226.40.18。

7 黄萱菁、夏迎矩、吴立德，基于向量空间模型的文本过滤系统[J]，软件学报，2003，14(3)：435-442。