

基于贝叶斯分类的邮件过滤方法及模型研究

肖 旻^{1,2}, 刘晓璐², 屠立忠²

(1. 东南大学 计算机科学与工程系, 江苏 南京 210096;
2. 南京工程学院 计算机工程系, 江苏 南京 210013)

[摘要] 垃圾邮件日益泛滥, 给用户带来了极大的不便和危害, 并对网络安全构成威胁. 传统邮件过滤方法单一, 过滤精度不高, 已不能很好地满足需求. 结合规则过滤技术, 分析了基于文本内容的贝叶斯分类器实现的关键技术与方法, 并给出核心过滤算法在邮件分类中的实现具体方法及过程, 进而完成垃圾邮件的判别. 为减少邮件的误判对用户造成的损害及垃圾邮件漏判造成的影响, 提出相应的改进措施, 使用最小风险贝叶斯决策减小误判率, 对分类系统经训练部分进行自适应调整, 最后给出基于规则与内容的双重防范机制的邮件过滤模型及基于该框架的邮件判别流程.

[关键词] 邮件过滤, 贝叶斯原理, 文本分类, 向量空间模型

[中图分类号] TP391 **[文献标识码]** A **[文章编号]** 1672-1292(2006)02-0086-04

Research in a Method and Model of Spam Filtering based on Bayesian Classifier

XIAO Min^{1,2}, LIU Xiaolu², TU Lizhong²

(1. Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China;
2. Department of Computer Engineering, Nanjing Institute of Technology, Nanjing 210013, China)

Abstract: The increasing junk mail brings great inconvenience and danger to people, threatens the safety of the network. The filtering way is single used by traditional filters, can't well satisfy the demand of filtering. This paper has analysed the key techniques and methods about Bayesian classifier of content-based, provided the effective way and process of kernelly arithmetic in filtering and completed the judgment of spam. In order to reducing the damages because of mistaking e-mail, we provide the improved methods of using the risk minimization Bayesian decision and self-improvement of categorization system. The paper finally has described a spam filtering model and process by double defending based on rule and content.

Key words: spam filter, Bayesian theory, text categorization, vector space model

0 引言

电子邮件已经成为人们日常生活中通信、交流的重要手段之一. 然而, 大量出现的垃圾邮件, 给用户造成时间和资源的浪费, 同时极大地消耗了网络传输资源以及邮件服务器的存储空间, 并对网络安全构成威胁. 针对这一问题尽快寻找解决方案的需求也更加迫切. 目前, 应对垃圾邮件的主要方法和手段有: 反垃圾邮件法制管理手段和利用邮件过滤技术进行处理. 前者可以依据立法, 对垃圾邮件制造者进行法律制裁, 但由于多种原因进展较慢; 后者在技术层面上解决垃圾邮件问题, 相继出现了多种邮件过滤技术. 目前, 各大邮件服务提供商或邮件客户端大多提供了一定的垃圾邮件过滤功能, 常用的包括黑名单与白名单技术、基于关键词搜索以及设定过滤规则等方法, 在实际使用中已逐渐不能满足过滤需求. 基于内容分析的文本分类技术正逐步进入邮件过滤技术当中. 本文在规则过滤的基础上, 结合基于文本分类以及贝叶斯 (Bayes) 理论的邮件过滤方法对垃圾邮件进行分类判别, 并提出进一步的改进措施, 给出邮件过滤模型.

收稿日期: 2005-09-28.

基金项目: 南京工程学院科研基金项目资助 (科研令号 04-37)

作者简介: 肖 旻 (1968-), 女, 讲师, 主要从事计算机应用及软件技术的教学与研究. E-mail: xiaomin_xy@sohu.com

1 关键技术

1.1 文本分类及表示

文本分类是文本处理领域的重要研究内容之一,是数据分析和模式识别中的一项基本任务.文本分类一般都由训练过程和分类过程两阶段构成.基于内容的文本分类通过一定数量的已分类好的训练文本,在学习各个类的训练文本的基础上,预测未知类文本的类别.目前在文本分类领域有多种分类方法,其中经典的方法主要有:K最近邻近法(k-Nearest Neighbor)^[1]、决策树法(Decision Trees)^[2]、贝叶斯分类法(Bayesian classifiers)^[3]、支持向量机(Support Vector Machine, SVM)^[4]方法等.

在信息处理领域,当前文本大多采用向量空间模型(Vector space model, VSM)表示,一篇文本可以表示为一个 n 维向量,即 $d(w_1, w_2, \dots, w_n)$,其中 w_i 为第 i 个特征项(Term)的权重, n 是特征项的个数,特征项可以是字、词、短语或者某种概念,本文采用词作为特征项.这样文本表示就转化为先进行文本分词,再由这些词作为向量的维数来表示文本,VSM中的权重计算可以采用基于数值型空间模型的词频TF(Term frequency,表示该特征词在文本中出现的次数)表示以及基于布尔型空间模型方法表示.TF-IDF(Term frequency-Inverse document frequency,倒排词频)表示方法目前较多用于数值型空间模型的权重计算.

1.2 特征项提取

由前述可知,用向量空间模型来表示文本时,由于向量空间的维数由文本集中词的数目来决定,因而维数是相当大的,然而文本的许多信息又是高度冗余的,所以需要进行降维处理工作.

本文进行向量维数压缩的主要思路和方法是:首先对文本进行预处理,去掉多次出现但不表现文本主题特征的词以及文本中出现频率过少的词,然后依据特征选择方法对词进行特征项选择,其间根据需要还可以添加其他特征,以提高分类效果.

1.3 邮件过滤中文本分类方法的选择

经典的文本分类方法以及机器学习理论大多可以应用于邮件过滤,无论对邮件服务器还是用户客户端,邮件过滤都对实时性要求比较高,因此要尽可能地采用计算简便、速度快的文本分类方法.而1.1节中基于概率统计的贝叶斯分类模型算法计算快捷,并能准确地对文本进行分类,且不需要花过多时间去训练样本集,运行速度较快,对于对文本数量、分类速度、过滤效率要求较高的邮件过滤应用来说,是一种较为理想的选择.结合邮件过滤特征,本文采用基于统计的朴素贝叶斯方法作为主要分类算法,并结合算法优化和过滤增强方式进行模型构建.

2 贝叶斯分类算法及实现

2.1 贝叶斯分类原理

贝叶斯分类原理最初源自于概率论中的贝叶斯定理.该定理表示对未来某件事情发生的概率可以通过计算它已经发生过的频率来估计.它在应用于邮件分类(分为垃圾和合法邮件)时,通过计算文本属于某个类别的概率,将该文本归为概率最大的类别中去,以判定邮件类别,在计算时,使用了贝叶斯定理的概率公式.

2.2 朴素贝叶斯模型及算法实现

朴素贝叶斯(Naïve bayes)^[3]分类模型是利用类别的先验概率和词的分布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.它建立在“贝叶斯假设”基础上,假定所有特征之间相互独立.应用到邮件分类中,它是通过邮件用户先提供一定数量的垃圾邮件和非垃圾邮件做为邮件训练集自动训练分类模型,将训练的结果作为判定未知邮件的主要依据,运用到相应分类算法中去.

由1.1节可知,每个文本都可以表示为一个 n 维特征向量 $d(w_1, w_2, \dots, w_n)$,对于给定的类变量 C_k ,基于其特征属性 w_1, w_2, \dots, w_n 之间相互独立假设,计算文本属于某个类别的概率 $P(C_k/d)$ 时,利用贝叶斯概率公式,对于给定的 d ,属于第 $C_k(k=1, 2, \dots, m)$ 类的概率为:

$$P(C_k/d) = P(C_k) \times P(d/C_k)/P(d) \quad (1)$$

$P(C_k)$ 是类的先验概率, $P(d/C_k)$ 是类条件概率.对同一个文本, $P(d)$ 不变.公式(1)中,后验概率 $P(C_k/d)$ 表示以该文本中出现的词与向量空间模型中特征项的匹配情况,决定该文本属于第 C_k 类的概率.可通过得到先验概率 $P(C_k)$ 和类条件概率 $P(d/C_k)$ 的值,来得到后验概率 $P(C_k/d)$.类条件概率可由

了合法邮件的误判率,改善了邮件分类过滤效果。

另外,垃圾邮件过滤中,由于垃圾邮件本身内容、形式的不断变化以及用户对垃圾邮件判别准则的改变,也容易造成垃圾的漏检问题,所以为用户提供自学习反馈机制,以适应不断变化的新情况。结合贝叶斯邮件过滤,可以选用过滤反馈学习方法中重新学习方式^[6],该方法将待学习的邮件集合和原来已学过的邮件集合组成一个新的学习集合,在这个集合上重新做一次特征选择,提取漏检邮件特征项,进行权值计算和分类器学习。与原有方式相比,提供自学习反馈机制扩展了特征项空间,增加了邮件训练集内容,增强模型自适应能力。

4 过滤模型结构

在实际应用中,需要将多种过滤方法结合起来,以达到较为理想的过滤效果。本文综合规则过滤和文本内容分类过滤两级模式进行模型构建。当规则过滤等手段不能给出确切的判断时,使用基于文本内容的贝叶斯分类过滤方法,进一步确定垃圾邮件。过滤模型及主要判别过程如图1所示。

Step 1 首先用户选取一定数量的邮件作为邮件训练样本,为分类决策提供依据,训练样本可以通过过滤系统进行自我学习不断扩展;

Step 2 对训练样本进行特征提取,提取出的内容放入内容特征库;

Step 3 新邮件到达时,经过邮件预处理后,经规则库中的规则匹配进行规则过滤,如果待检邮件被判为垃圾邮件,则直接按垃圾邮件处理;否则,待检邮件进入第二级过滤,转到Step 4;

Step 4 按文本内容的贝叶斯分类进行过滤,利用内容特征知识库及贝叶斯算法进行文本分类与决策区,判别垃圾邮件与合法邮件,分别进行邮件处理;

Step 5 最后过滤系统经反馈邮件分类决策结果到邮件训练部分进行自学习,动态更新特征库,增强过滤系统自适应能力。

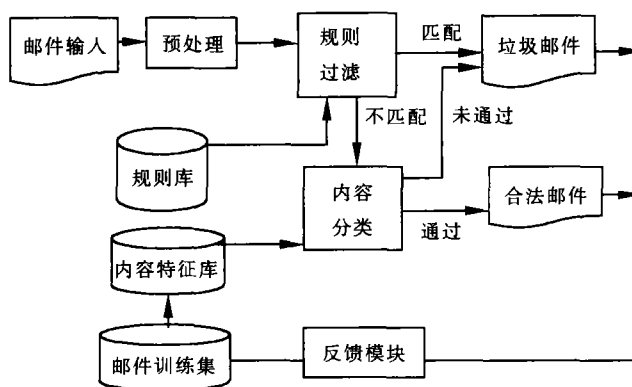


图1 邮件过滤模型框图

5 结语

本文讨论了基于内容的贝叶斯分类的邮件过滤方法和过滤模型,提出的解决方案在克服传统垃圾邮件过滤手段单一、过滤精度低、缺乏智能性等缺陷方面有所改进,但在综合过滤技术解决方案中,也有不利影响,例如对于邮件的准确率和查全率等指标还需要做进一步提高,需对算法继续进行调整。

[参考文献] (References)

- [1] YANG Yiming. A example-based mapping method for text categorization and retrieval[J]. ACM Transactions on Information Systems, 1994,12(3):252-277.
- [2] CARRERAS X, MARQUE L. Boosting trees for anti-spam email filtering[C]//Proceedings of Euro Conference Recent Advances in NLP(RANLP-2001). [S.l.]:[s.n.], 2001:58-64.
- [3] MEHRAN S, SUSAN D, DAVID H, et al. A bayesian approach to filtering junk E-mail[C]//Proc of AAAI Workshop on Learning for Text Categorization. Madison, Wisconsin, 1998: 55-62.
- [4] DRUCKER H, VAPNIK V. Support vector machines for spam categorization[J]. IEEE Transactions On Neural Networks, 1999,20(5):1048-1054.
- [5] LIN Yaping, CHEN Zhiping, YANG Xiaolin, et al. Mail filtering based on the risk minimization Bayesian algorithm[J]. Proceedings Industrial System and Engineering, 2002,17(3):282-285.
- [6] 王斌,潘文峰. 基于内容的垃圾邮件过滤综述[J]. 中文信息学报, 2005,19(5):1-10.
WANG Bin, PAN Wenfeng. A survey of content-based anti-spam email filtering[J]. Journal of Chinese Information Processing, 2005,19(5):1-10. (in Chinese)

[责任编辑:刘 健]