基于朴素贝叶斯的垃圾邮件分类系统的设计

徐治国
(盐城民航站,江苏 盐城 224051)

摘要:结合垃圾邮件分类系统的具体要求,在传统规则分类方法的基础上引入机器学习的知识,给出了系统体系结构和特征提取算法,试验了一种对新邮件计算所属类别后验概率的方法,并详细讨论了一个基于朴素贝叶斯方法的个性化垃圾邮件分类系统的设计。提出的分TFIDF特征子集提取算法和朴素贝叶斯方法对邮件进行分类具有较好的分类精度,应用朴素贝叶斯方法在新邮件到达的同时对其进行分类,具有较好的分类速度。

关键词:电子邮件;文本分类;朴素贝叶斯;机器学习
中图分类号:TP39　　文献标识码:A　　文章编号:1671-5322(2008)02-0047-04

由于信息技术特别是Internet的发展和E-mail的普及应用,各种文本信息急剧增加,文本分类成为处理和组织大规模文本信息的关键技术。

文本分类是指在给定的分类体系下,根据文本的内容确定文本相关类别的过程。文本分类的方法有决策树分类方法、K—最临近分类方法、朴素贝叶斯分类方法和KNN算法等。不同算法的精度各不相同,适用的领域也不一样。在这些算法中,朴素贝叶斯分类方法的结果比想象中要好。文本分类的关键问题是如何构造一个分类函数或分类模型(也称为分类器),并利用此分类模型将未知文本影射到给定的类别空间。分类器的构造方法有多种,主要有统计方法、机器学习方法、神经网络方法等。国外对文本分类技术的研究已经开展了多年,并在邮件分类、电子会议、信息过滤等方面得到了较为广泛的应用。

随着E-mail的日益普及,我们注意到网络管理面临的新问题垃圾邮件的泛滥。根据伦敦的电脑安全防护公司MI2G在一份报告中指出,垃圾邮件在2003年10月给全球经济造成104亿美元的损失,远远超过了电脑病毒和黑客造成的损失。因此对垃圾邮件进行有效的分类,是一个有重要意义的现实问题。

1　分类系统体系结构和关键技术

1.1　垃圾邮件分类基本过程

图1所示为一个基于朴素贝叶斯分类方法的垃圾邮件分类系统的结构框图。它实质上是一个

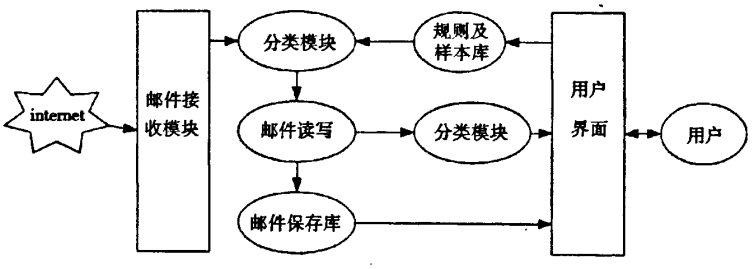

图1　基于文本分类的垃圾邮件分类系统的结构框图

Fig. 1　Structure figure of junk mail classification system based on text categerization

收稿日期:2008-03-12

作者简介:徐治国(1973-),男,江苏盐城市人,工程师,主要研究方向为电子信息通讯。

多类别的文本分类器,它以 E—mail 作为分类对象,通过从给定的不同类别的邮件训练例集中提取不同类别对应的特征模式,再以此为基础,对新邮件用朴素贝叶斯的方法进行分类,计算出新邮件属于和不属于各类别的后验概率,从而确定它最有可能属于的那个类别。"个性化"体现在用户可以各自定义自己的分类类别。

1.2 文本预处理

使用最普遍的文档表示方法是向量空间模型(VSM:Vector Space Module)采用 VSM 要求对文本进行处理,以得到文本的向量表示。文本预处理通常与语言、文本格式及拟采用的分类算法都是相关的。

采用 VSM 的一般步骤是:先通过对训练语料的学习对每个类建立特征向量作为该类的表征,然后依次计算该向量和各个类的特性向量的距离,选取距离大小符合阈值的类别作为该文本所属的最终类别。

在系统中,采用了两个模块来实现垃圾邮件的分类系统,分别为训练模块和识别模块,分别对应文本分类的训练和识别这2个阶段。工作流程如图2所示:

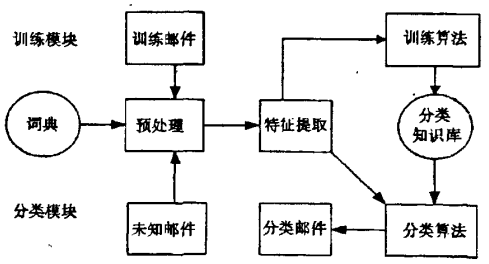

图2 垃圾邮件分类系统工作流程

Fig. 2 Flow chart of junk mail classification system

训练模块是对训练邮件进行预处理、特征选择和提取、参数训练,生成分类知识库;识别模块首先对要识别的邮件进行预处理、特征提取,然后用训练得到的分类知识库通过识别算法对邮件进行自动分类。

1.3 特征提取

由于特征提取中的基本元素是词条,特征向量的维数通常是惊人的,而且包含了很多噪声,因此需要通过特征约减删除提取后的文本向量中对文本分类没有显著贡献的特征项。考虑提高效率和除去噪音的目的,在文档表示为可用于分类的表示形式之前,需要进行特征选择。特征选择是从每一类文档的所有特征中抽取那些能够反映和区分此类文档与其它类文档的特征项,这是分类问题的关键。文本分类中的特征选择一般是通过大量已知类属的文本,统计出能够反映此类文档的特性。特征选择常采用的处理方法包括:TFIDF、互信息、词频、信息等。

在本垃圾邮件分类系统中,考虑到邮件的处理都在服务器端操作,对系统速度和效率要求比较高,故需要一种简单易懂,处理速度快的算法。而 TFIDF 满足这种要求,故采用 TFIDF 方法来进行特征提取。具体操作如下:

邮件是一种半结构化的文本,一封邮件由邮件头(结构化信息)和邮件体(无结构化信息)组成。前者包括:From、Sender、Send Time、Subject 等信息,后者指的是邮件正文。显然单纯根据结构化信息或只根据邮件体对邮件进行分类都是有所偏颇的,毕竟两者中都包含了对分类极为重要的信息。在本系统中,我们对每一封训练例子不仅从头部信息中提取特征,同时也从邮件体中提取特征,共同构成这封邮件的特征模式。

在系统实现过程中,主要采用下面几项来表示特征:特征名称、在所有类别的训练例中包含该特征的文档个数、在本类别的训练例中包含该特征的文档个数、该特征的权重等。其中,特征名称表示对邮件正文或主题部分进行挖掘所得到的一些词或词组。

本垃圾邮件处理系统框架见图3:

1.4 朴素贝叶斯分类器的设计

朴素贝叶斯分类器是一种简单而有效的分类器,实际使用表现出了很好的分类准确率和速度,它假设所有属性都条件独立于类变量 C。朴素贝叶斯分类的做法如下:

设每个数据样本用一个 n 维向量 $X=\{x_1, x_2, \cdots x_n\}$,分别描述对 n 个属性 $A_1, A_2, \cdots, A_n$ 的度量。通过概率计算,根据待分类实例的向量 $X'=\{x'_1, x'_2, \cdots, x'_n\}$,求出最可能的分类目标值。即对于每一个 $C_i \in C$,计算条件概率 $P(C_i/x'_1, x'_2, \cdots, x'_n)$,并取 $\max\limits_{C_i \in XC}\{P(C_i/x'_1, x'_2, \cdots, x'_n)\}$ 对应的类别作为目标输出值。根据贝叶斯定理,

$$P(C_i/x'_1, x'_2, \cdots, x'_n) = [P(x'_1, x'_2, \cdots, x'_n/C_i) * P(C_i)]/P(x'_1, x'_2, \cdots, x'_n) = \prod_{j=1}^{n} P(x'_1/C_i) * P$$

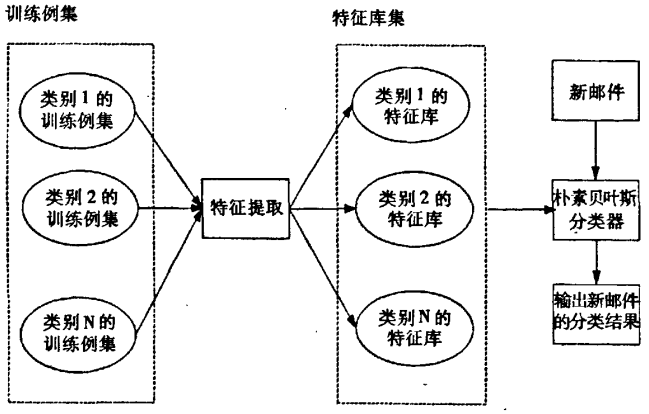

图 3　垃圾邮件处理系统框架图

Fig. 3　Structure figure of junk mail processing system

$(C_j)/P(x'_1,x'_2,\cdots,x'_n)$

$P(x'_1/C_i)$表示特征 x'_i 在类别 C_i 对应的训练例集中出现的概率$(j=1,2,\cdots\cdots,n)$。

在上述方法中,要求每一个数据样本用统一的属性集$\{A_1,A_2,\cdots,A_n\}$来刻画,而在本垃圾邮件分类系统中,由于从类别 C_i 的训练例集中提取的特征子集 C_iF 从类别 C_j 的训练例集中提取得到的特征子集 C_jF 的维数通常是不相同$(i\neq j)$,所以无法直接使用上述方法,因此作者使用了如下设计的贝叶斯分类器。

假设 $C=\{C_1,C_2,\cdots,C_k\}$,其中 $C_i(i=1,2,\cdots,K)$表示第 i 个类别,本文用 C_iF 表示第 i 个类别的特征子集,设 $C_iF=\{Cf_1,Cf_2,\cdots,Cf_N\}$,$Cf_j$ 表示 C_i 的第 j 个特征 f_j。对于一封新邮件 e,用所有的 $C_iF(i=1,2,\cdots,K)$分别对 e 进行"特征化",得到一组向量集 $X_e=\{X_1,X_2,\cdots,X_k\}$。其中 $X_i(i=1,2,\cdots,K)$为用 C_iF"特征化"e 得到的向量,设 $X_i=\{x_{i1},x_{i2},\cdots,x_iN\}$,如果 $Cf_j(j=1,2,\cdots,K)$在 e 中出现,则 x_{ij}为 1,否则为 0。对每一个 X_i,分别计算 $P(C_k/X_i)$和 $P(C_k/X_i)(i=1,2,\cdots,K;k=1,2,\cdots,K)$。

其中,$P(x_{ij}/C_k)$的计算方法为:如果 x_{ij}为 1,则 $P(x_{ij}/C_k)$表示在 C_k 类的训练例集中包含特征 C_kf_j 的例子占该类训练例总数的比例;如果 x_{ij}为 0,则 $P(x_{ij}/C_k)$表示在 C_k 类的训练例集中不包含特征 C_kf_j 的例子占该类训练例总数的比例;$P(C_k)=C_k$ 类的训练例总数/所有类别的训练例总数。同理,可以计算 $P(C_k/X_i)$,但由于在 C_k 类的例子中 C_kf_j 可能不会出现,即存在 j,使 $P(x_{ij}/C_k$ 的值为 0,从而导致 $P(C_k/X_i)$的值为 0,此时可用一个正小数来替代 $P(C_k/X_i)$,本垃圾邮件分类系统中采用的是 0.01。最后,取 $\max\limits_{C_k\in C}[P(C_k/X_i)+P(C_k/X_i)]$对应的那个类别作为新邮件对应的类别。

1.5　阈值确定

不同文档类所对应的特征向量和相关阈值也不相同。上述的算法方案都是将新邮件归于分类体系中的一个类,即与该邮件关联最大的类,而事实上,分类体系中的类别不是完全互斥的,存在这样一些既属于其中一个类别,又同时属于其他类别的邮件,对于这种邮件,上述的分类算法无法确定邮件所属的所有类别。因此,需要对每个类别确定阈值,当邮件在该类的阈值之上时,就将邮件归于该类中。

在系统中考虑了一种确定阈值的方法,称为百分比阈值确定法,它的基本思想是:首先依据训练算法和分类算法构造分类器,然后对于要确定阈值的类,用分类器分类该类中所有的训练邮件,从而每个邮件都得到一个相关的值。以上述算法为例:

贝叶斯算法:邮件归属于类的几率,然后按递减顺序排列所有本类训练邮件得到的值,假定本类有 n 篇邮件,那么这些邮件的值为 $d_1,d_2\cdots\cdots d_n$,那么本类阈值 y 确定如下:

$$y=d_{sn\%}$$

其中,s 为初始值,根据训练邮件的质量程度,可以确定为 80 或更高,这样就确定了本类的初始阈值,可以想象,S 越大,该分类器的查全率就越高,

准确度就越低,相反地,S 越小,查全率就越低,准确率就越高,然后根据测试进行调整。

相应地,调整阈值可以转化为调整 s 值,如果对查全率满意而对准确率不满意,那么可以减少 s 值,否则就增加 s 值。

2 系统评估

因为文本分类从根本上说是一个映射过程,所以评估文本分类系统的标志是映射的准确程度和映射的速度。映射的速度取决于映射规则的复杂程度,而评估映射准确程度的参照物是通过专家思考判断后对文本的分类结果,与人工分类结果越相近,分类的准确程度就越高,评估文本分类系统的主要有两个指标:准确率和查全率。

准确率是所有判断的文本中与人工分类结果吻合的文本所占的比率。

其数学公式表示如下:

$$\text{准确率(precision)} = \frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}$$

查全率是人工分类结果应有的文本中分类系统吻合的文本所占的比率,

其数学公式表示如下:

$$\text{查全率(recall)} = \frac{\text{分类的正确文本数}}{\text{应用文本数}}$$

准确率和查全率反映了分类质量的两个不同方面,两者必须综合考虑,不可偏废,因此,存在一种新的评估指标,$F1$ 测试值,其数学公式如下:

$$F1\text{ 测试值} = \frac{\text{准确率} \times \text{查全率} \times 2}{\text{准确率} + \text{查全率}}$$

3 结束语

本文对朴素贝叶斯方法用于垃圾邮件的处理进行了探讨,并设计了垃圾邮件的分类器,提出了百分比阈值确定法,并对系统进行了评估,认为具有较好的分类精度和分类速度,取得了阶段性的成果。我们将在算法优化和系统实现等细节方面作进一步的研究。

随着互联网和多媒体技术的进一步发展,邮件分类技术将与图像识别、语音识别融合,比如图像邮件的分类、语音邮件的分类,多媒体数据库索引等。这就进一步要求邮件分类技术在文本的处理方法、克服噪音干扰、分类精度方面有进一步的提高。总之,邮件分类技术也会是将来研究的热点问题,它的发展同时会推动数据库技术,网络技术的共同进步。

参考文献:

[1] Saiton G, McGill M J. An Introduction to Modern Information Retrieval. New York: McGraw - Hill Book Co, 1983.
[2] 刘壁松,李春平. 一个可扩展的文本分类系统的设计与实现[J]. 计算机工程与应用,2006(4):25 - 27.
[3] 邹涛. 中文文档自动分类系统的设计与实现[J]. 中文信息学报,2006(3):68 - 71.
[4] 彭树青. Internet 垃圾邮件过滤技术研究[J]. 信息技术,计算机科学,2006(6):81 - 83.

The Design of Junk Mail Classification System Based on Na? ve Bayes

XU Zhi-guo
(Yancheng Civil Aviation Station, Jiangsu Yancheng 224051, China)

Abstract: The research of anti junk mail is the hotspot in computer science research area at all times。 This paper combines the specific demand to junk mail classifier, introduces the knowledge of machine learning on the base of the traditional regular classification, presents the architecture of the junk mail system and the feature extraction algorithm, and tests a new method to compute the posteriori probability which sort a new email fall into, and discusses in detail the design of an individual junk mail classifier which is based on Na? ve - Bayes. When the system uses the dispart words algorithm, TFIDF feature subset abstraction algorithm and Naive - Bayes method, it classifies emails more precisely and more quickly.

Keywords: email; text classification; Na? ve - Bayes; machine learning