第八届“泰迪杯”数据挖掘挑战赛

题　目：“智慧政务”中的文本挖掘应用

关键词：文本过滤；朴素贝叶斯；垃圾词库；Bayes;SVM—EM

摘要

互联网技术的快速发展为社会网络互动平台提供了强有力的支持，由于留言板被关注度高，是广大人民等关心社会发展、反馈疑难问题的重要渠道，是社会管理的重要辅助工具，因此留言板在信息安全管理中不可忽视。社区网留言板是上网经常浏览发布信息的平台，因此保证留言板网络环境的纯净。根据社区的实际情况，让网络管理员以最少的工作量简单易行的检测过滤出垃圾留言是本文研究的主要内容。

本文将过滤留言板内的留言作为过滤研究目标。提取某地留言板中网页显示留言和管理后台被屏蔽留言做为样本。 在研究过程中主要做出以下工作重点：

（1）利用朴素贝叶斯分类算法，将文本分类技术应用留言板留言过滤中。计算留言在不同类别中的概率，选择出概率较大的类别，完成对留言的分类，实现留言过滤。由于数据样本较少，为了更好验证算法效果，采用 3-折交叉验证法进行朴素贝叶斯过滤实验。

（2）提出基于规则与朴素贝叶斯相结合的过滤模型。 朴素贝叶斯分类结果中会出现一些留言被漏判的情况。为了提高留言过滤的召回率，提

出对贝叶斯分类结果进行二次过滤。通过人工建立词库，将贝叶斯分类结果进行字符串的匹配，从而实现对漏判留言的修正。

Bayes 算法在已知先验概率与条件概率的情况下进行模式分类,待分样本的分类结果取决于各类域中样本的全体,但实际上类别总体的概率分布和各类样本的概率分布函数是不确定的。为了解决上述问题,提出了一种基于 SVM-EM 算法的 Bayes 算法,首先利用非线性变换和结构风险最小化原则将流量分类问题转化为二次寻优问题,然后要求 EM 算法对 Bayes 算法要求条件独立性假设进行填补,最后利用 Bayes 算法进行网络留言分类,提高了分类的准确性和稳定性。

Abstract

The rapid development of Internet technology provides a strong support for the social network interactive platform. Because message board is highly concerned, it is an important channel for people to care about social development and feedback difficult problems, and an important auxiliary tool for social management. Therefore, message board cannot be ignored in information security management. The community network message board is the Internet often browse the release of information platform, so to ensure the message board network environment of the pure. According to the actual situation of the community, the main content of this paper is to let the network administrator to detect and filter the spam message with the least amount of work.

This paper will filter the message board as the target of filtering. Extract a certain message board in the web page display message and management background is blocked as a sample message. In the research process, the following work focuses are mainly made:

(1) apply text classification technology to message board message filtering by using naive bayes classification algorithm. Calculate the probability of message in different categories, select the category with greater probability, complete the classification of message, and realize message filtering. Due to the small number of data samples, in order to better verify the effectiveness of the algorithm, 3-
fold cross validation method was used to carry out the naive bayesian filtering experiment.

(2) proposed a filtering model based on rules and naive bayes. In the naive bayesian classification results, some comments will be missed. In order to improve the recall rate of message filtering, secondary filtering of bayesian classification results was proposed. The bayesian classification results are matched by strings through the artificial word base, so as to correct the message of omission.

Bayes algorithm conducts pattern classification under the condition of known prior probability and conditional probability, and the classification result of samples to be divided depends on all samples in all kinds of domains, but actually the probability distribution of category population and the probability distribution function of all kinds of samples are uncertain. In order to solve the above problems, this paper proposes a Bayes algorithm based

on SVM - EM algorithm, first of all, using nonlinear transformation and structural risk minimization principle will flow classification problem into a quadratic optimization problem, and then ask

EM algorithm to fill the Bayes algorithm requires conditional independence assumption, finally use the Bayes algorithm network message classification, improves the classification accuracy and stability.

1.问题重述

1.1 问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依 靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、 云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。 附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的 答复意见。请利用自然语言处理和文本挖掘的方法解决下面的问题。

1.2 研究背景及意义

因为时代的发展，计算机技术的不断进步和互联网的蓬勃发展给人们的生活带来了翻天地覆的变化。普通人通过互联网不仅增加了信息获取途径，同时也丰富了大家娱乐生活，成为广大网民交流沟通的重要平台，同时也促进了经济的发展和社会的进步。

由阮彤、冯东雷、李京三人共同提出基于贝叶斯网络的信息过滤模型BMIF。该模型利用图表直观表示整个网络，利于理解方便交互；且能同时布尔模型与向量模型，将合作过滤与基于内容过滤方法结合起来。上海交通大学信息安全工程学院的丁霄云等人，通过使用特征简约，对向量机的维数进行约束，实现了对 SVM 算法改进，有效减少了 SVM 的维数。 通过互联网调查众多中职学校的留言板，不难发现，在实际生活应用里中职学校留言

板上的垃圾信息大多是一些广告、暴力语言、愤怒发泄留言等。针对这种情况，学校可以基于现有的服务器等网络资源，在留言板的管理后台将留言信息进行提取分类，删除垃圾留言，发布合法留言，即可实现对留言板垃圾留言的过滤。

2.模型假设

（1）假设记录仪器在记录过程中因环境因素导致的数据误差可以忽略。

（2）假设车辆行驶过程中的因车辆硬件问题导致的数据误差可以忽略。

（3）假设问题二中每辆车的所有不良行为是由驾驶员导致的。

（4）假设附件 1、2 中的数据完全真实，不存在误差。

（5）假设研究附件 1、2 中的这些部分数据，可以作为整体考虑。

3.模型的建立与求解

3.1 问题一的建立与求解

3.1.1 问题分析

文本分类(Text Categorization)是指依据文本的内容,由计算机根据某种自动分类算法,把文本判分为预先定义好的类别。文本分类是信息存储和信息检索中的重要课题。互联网的飞速发展又给文本分类提供了新的应用平台。网页分类是文本分类在网页文本集合上的应用,它在信息过滤,基于个性化的信息服务等方面有着重要用途。网页自动分类具有如下优点:不需要人工干预,节省大量人力物力,更新快,而且分类速度较快,精度较高,满足实际应用要求。

文本分类大致可分为三个步骤:文本的向量模型表示,文本特征选择和分类器训练。数量巨大的训练样本和过高的向量维数是文本分类的两大特点。这两个特性决定了文本分类问题是一个运算时间和空间复杂度很高的学习问题。为了兼顾运算时间和分类精度两个方面,我们不得不进行特征选择,力求在不损伤分类性能的同时达到降维的目的。

在文本分类中,常用的特征选择方法有基于阈值的统计方法,如文档频率方法(DF)[2],信息增益方法(IG)[2],互信息方法(MI)[2], CHI[2]方法,期望交叉熵[3],文本证据权[3],优势率[3],基于词频覆盖度[4]的特征选择方法等,以及由原始的低级特征(比如词)经过某种变换构建正交空间中的新特征的方法,如主分量分析[5]的方法等。基于阈值的统计方法具有计算复杂度低,速度快的优点,尤其适合做文本分类中的特征选择,在本文中将集中研究和比较 8 种基于阈值的统计方法。关于文本分类中的特征选择问题,比较有代表性的是 Yang Yiming[2]和 DunjaMladenic[3]的工作。前者针对平面文本分类问题,分析和比较了 DF,IG,MI 和 CHI 等 5 种方法,结合 LLSF 和 KNN 分类器,得出 IG 和 CHI 方法效果相对较好的结论。而后者针对等级文本分类问题,分析和比较了信息增益,期望交叉熵,文本证据权及优势率等方法,结合 Naïve Bayes 分类器,实验结果表明二元优势率是最好的选择方法。由于两者针对不同类型的分类问题,不同的实验数据集,采用不同的分类器,因此得出了不太一致的实验结论。为了综合研究各种特征选择方法的选择性能,在相同的平面文本分类问题上对各种选择方法进行实验和比较是必要的。

在分类算法的选择上,目前存在各种各样的文本分类算法,如文本相似度法[6](也称向量空间法),Naïve Bayes 方法[3,6], K-最近邻算法[7](K-Nearest Neighbor),Neural Network 方法[8],SVM 方法[9]等。文本相似度方法和 Naïve Bayes 方法是应用最多的两种方法,它们具有分类机制简单,处理速度快的优点。

在前人工作基础上,我们不仅研究和比较了信息增益,期望交叉熵等六种常用的特征选择方法,而且把应用于二元分类器中的优势率改造成适用于多类问题的形式,并提出了一种新的类别区分词的特征选择方法。根据词的类间后验概率分布进行区分性定义,把某个词最可能出现的那一类和其他类别区分开来,这种区分性越大,那么该词就越可能是某一类的核心特征,由此选出那些强类别意义的分类特征。结合文本相似度方法和 Naïve Bayes 分类器,在两个类别样本分布不同的网页集上作训练和测试,结果表明:改造的多类别优势率和类别区分词的方法取得了最好的特征选择效果。

3.1.2 问题求解

信息过滤模型

由于互联网发展迅猛，用户面对网络上的海量信息需要进行甄别筛选。根据用户在工作、生活中遇到的实际情况，产生了对信息筛选的不同需求。用户需求可以是用户想要保留的信息内容，同样也可以是用户不想保留的信息内容。可以利用匹配算法进行匹配，在所有的信息中正确选择出用户需要保留下的信息，忽略用户不需要保留的信息，实现过滤的目

标。因为需要涉及过滤的成功率，所以用户需要收到过滤系统的反馈信息，从而不断完善自己的需求表示，使筛选信息的目标更明确，以求达到最好的过滤效果。 信息过滤的原理可以用图简单模型进行表示。

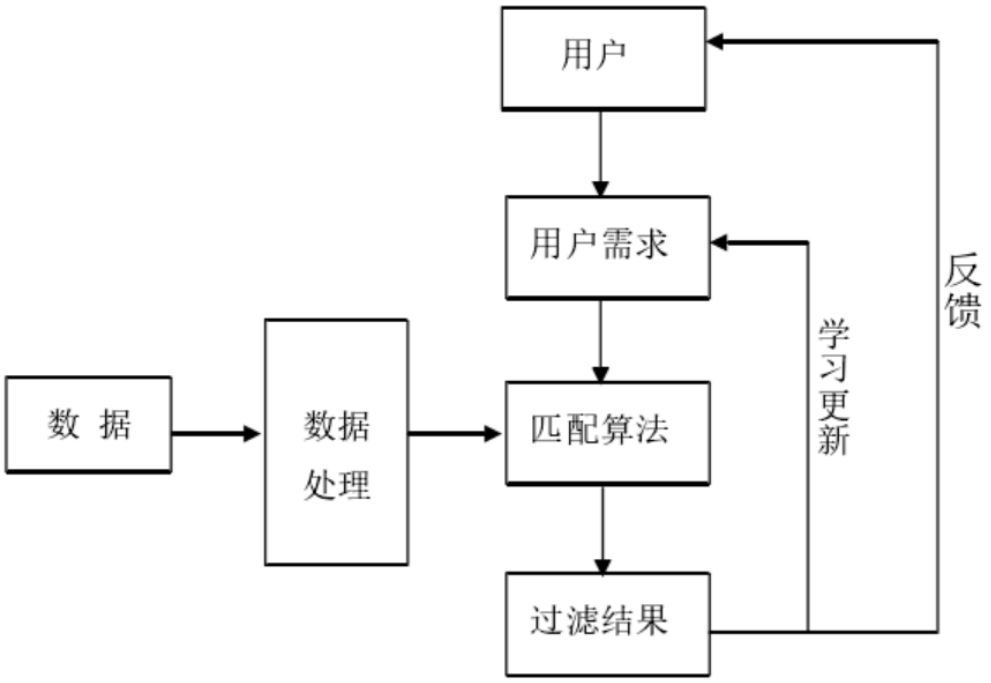
通过图中的信息过滤模型可以看出，信息过滤的核心内容可以表示为：

第一，用户需求的表述；

第二，将待处理的信息和用户要求条件进行匹配；

第三，过滤结果对用户进行反馈；

第四，用户需求的更新。



2.1.信息过滤种类及方法

二、信息过滤方法

信息过滤方法是多样化的，考虑角度不同，衍生出的过滤方法也是不相同的：根据对网络信息是否进行预处理分类，信息过滤分为主动过滤和被动过滤；根据过滤目标的不同，信息过滤又可分为内容过滤、网址过滤及混合过滤三种。

常见的不良信息过滤方法有以下几种：

1、分级法

分级法就是以网页内容或者其他特征为依据，将网络信息按照不同特性以用户需求为标准进标记分级，由过滤模板进行判断，实现过滤。网页制作者、网页使用者或者第三方可以是进行网页分级的操作者。自我分级的操作者是由网页制作者本人进行的，而第三方分级的操作者则是由网页使用者或者第三方进行的。

自我分级法因为是人工进行分级，所以管理方便，过滤错误率较低，但是分级质量无法保证。第三方分级的数据库庞大，分级成本随着分类项目的增多而增多，分级标准无法及时和用户沟通，过滤的效果不甚理想。

2、URL 地址列表法

在所有法信息过滤的方法中最直接的是 URL 地址列表法。在过滤前需要提前收集和建立完成 URL 地址表，并根据这个地址表决定用户允许或者禁止访问那些网址。这种方法简单直接，列表中的地址只有两种：白名单和黑名单。白名单（white lists）就意味着列表内的 URL 地址是允许被访问的，相反的黑名单（black lists）中的 URL 地址则是禁止被访问的。

URL 地址列表法容易实现，利用白名单可以限定访问范围，但访问地址的覆盖面很难满足用户需求。黑名单内被禁止访问的地址收集和维护比较麻烦，不能随着用户需求的变化而变化，无法满足用户信息需求的多样化。

3、动态文本分析法

前面介绍的两种过滤方法，既简单又容易实现，不过因为他们都无法实现在过滤的过程中根据用户需求的改变自动更新，从而影响了过滤效果。动态文本分析法可以解决这种问题，实现动态过滤。采用动态文本分析法进行过滤的过程中，首先是由用户提出过滤需求，然后描述出的用户需求模板必须满足用户实际需求。完成过滤的过程就是把动态文本与用户需求模板进行匹配，最后还需要利用反馈信息对用户需求模板进行改进，以达到良好的过滤效果。

在信息过滤的过程中，采用动态文本分析方法进行过滤时，有四个关键问题需要重点处理：建立用户需求模板，提取动态文本信息，匹配算法的选择，过滤结果的反馈。

文本分类算法

信息过滤是要判断信息是否符合用户的需求，如果符合需求则被保留，反之则被过滤。因此从本质上说，信息过滤可以看做一个“是”与“否”的二类分类问题，所以文本分类技术大多也在文本信息过滤中使用。在文本信息过滤的实际操作中，可以将根据用户需求筛选文本的过程看作是一个分类问题，即将文本分为符合用户需求类和不符合用户需求类，不符合用户需求的类别就被过滤掉。在信息过滤模型中，因为信息的过滤涉及实时性问题，所以分类算法需要选择那些简单快捷的算法，同时还要选择准确率高的算法，以避免符合用户需求信息的误判。比较常用的文本分类算法有：支持向量机法、决策树算法、贝叶斯算法、K 邻近法。

一、支持向量机法

支持向量机法可以表示为 Support Vector Machines（即 SVM），由科尔斯特（Cortes）和瓦普尼克（Vapnik）于 1995 年首先提出。SVM 其实是一种二分类模型，它的基本模型是定义在特征空间上的，是一种间隔最大的现行分类器。其不同于其它系统的优点主要表现在解决小样本、非线性及高维模式识别等问题上面。支持向量机的中心思路是：将多个待分类的点映射到“高位空间”，接着在这个新的高位空间中找到一个可以将这些点分开的“超平面”（最优线性分类面），同时要找到这两类点之间的“最大间隔”。

2　基于 SVM-EM 的朴素贝叶斯分类算法

朴素贝叶斯算法是一种简单而高效的分类算法,但是它的条件独立性假设极大影响了分类性能,因此,笔者提出了一种基于支持向量机 SVM、EM 算法融合的改进朴素贝叶斯分类算法: SVM-EM-NB 算法,该算法通过贝叶斯统计方法对网络留言进行分类,挖掘有效的数据信息,并结合支持 SVM 技术预测垃圾留言概率,贝叶斯前验分布和后验分布用来估计 SVM 中的参数。

2.1　支持向量机 SVM 训练

支持向量机 SVM 是目前一种新兴的技术,在文本分类方面越来越受到重视。支持向量机 SVM 的提出有很深的理论背景,其训练的本质是解决一个二次规划问题,得到的是全局最优解。

SVM 训练的基本思想概括为:

1)针对线性可分情况进行分析,通过使用核函数与非线性转换算法,将低维输入空间线性不可分的样本变化为高维特征空间并使其线性可分,从

而使得高维特征空间可以采用线性算法对样本的非线性特征进行线性分析。

2)基于结构风险最小化理论在特征空间中构建最优分割超平面,使得学习器得到全局最优化。 SVM 分类器的优点在于通用性较好,可以提高泛化性能,解决非线性问题,且分类精度高,分类速度与训练样本个数无关,其和朴素贝叶斯分类算法融合将大大提高查准率和查全率。

设分类线性方程为 x• w+b=0,对它进行归一化,使得对线性可分的样本集(xi,yi)(i=1,… ,n,n∈ Rd,y∈ {+ 1,-1})满足约束条件 yi[(w• xi)+b]-1≥ 0(i=1,… ,n)。利用 Lagrange 优化方法将最优分类面问题转化为对偶形式,即:在约束条件∑yiTi=0 和 Ti≥ 0,i=1,… ,n 下,对 Ti 求解下列函数的最大值 MaxQ(T)=∑Ti-∑TiTjyiyj(xi xj),其中,Ti 为原问题中与每个约束条件相对应的 Lagrange 乘子。这样就转换为一个不等式约束下二次函数寻优的问题,即存在唯一解。对应的样本是支持向量,采用线性分类解上述问题后得到的最优分类函数为 f(x)= sgn{(w x)+ b}= sgn∑Tiyi(xi x)+ b ,其中,b 表示分类阈值,可以用任一个支持向量求得。

对非线性问题,通过非线性映射 H 把输入空间的样本转化为某个高维特征空间,核函数 K(xi,xj)=H(xi)•H(xj),在高维空间进行内积运算时,无需了解变换 H 的形式。这样不仅能实现非线性变换后的线性分类,而且还没有增加时间复杂度,此时对偶形式变为 MaxQ(T)=∑Ti-∑TiTjyiyjK(xi xj),而相应的分类函数变为 f(x)= sgn∑TiyiK(xi,x)+ b。

SVM 将大的分类工作量放在输入空间而不是高维特征空间中完成,避免算法可能导致的“维数灾难”,快速解决二次规划的问题,具有较高的训练准确度。

2.3 SVM-EM-NB 算法

首先用优化训练的 SVM 训练留言集解决一个二次规划问题,使学习器得到一个全局最优解,然后把数据集分成完整集和缺失集,计算缺失属性的数据项与完整属性数据项的相关度,取相关度最大的数据项对应的属性作为缺失属性的一个估计值,此估计值作为 EM 算法的初始值,然后执行 EM 算法的两步,完成极大似然估计,用最后估计的值来完成缺失属性的填补,最后用朴素贝叶斯分类算法对完整数据集进行分类。

输入: T={X1,X2,… ,Xn},其中,X1,X2,… ,Xn 为原始属性集, λ={t1,t2,… ,tm}为类别属性。

输出:样本 X 的类别。

算法主要步骤为:

步骤 1 把数据集 T 分为 2 个数据子集 Ti 和 Tj。Ti 中的记录全部为完整记录,任何属性不含缺失值;Tj 中的记录为不完整记录,即属性中含有一个及以上的缺失值。

步骤 2 调用 EM 算法,完成缺失数据填补。

步骤 3 随机选择 4/5 的样本作为训练集,剩余 1/5 的样本作为测试集,计算训练集样本的先验概率 P(λ)。

步骤 4 在假设类条件独立的情况下,根据贝叶斯公式计算条件概率 P(Xi|λ)。

步骤 5　根据式(1)计算后验概率 P(λ｜ Xi),输出类别,求出分类准确率。

向量空间模型在文本分类领域,最常用的文本表示模型是 G..Salton 在 1975 年提出的向量空间模型(VectorSpaceModel),其基本思想是把文本 di 看作向量空间中的一个 n 维向量(ti1,w(ti1),ti2,w(ti2),… tin,w(tin)),其中 ti1,ti2,… ,tin 为表示该文本的 n 个特征,w(tik),k=1,2,… ,n 是该文本对应第 k 个特征的权重,一般取为词频的函数。对于中文文本来说,由于词是语义的最小单位,因此一般选择词作为特征。各维特征通常表示成词频 tf(tk)和反文档频率 idf(tk)的函数,即有:w(tik)= tf(tik)× idf(tik)。其中 tf(tik)表示词 tk 在第 i 篇文档中出现的次数,而 idf(tik)= log(N df(tk)),N 为文档集中的全部文档数,而 df(tk)表示出现词 tk 的文档数。为了计算方便,通常还要对向量进行归一化。

作为网页分类的第一步,我们对中文网页集进行基于词典的分词处理,由于所选用的通用词典共有 116921 个词条,因此把每个网页表示为 116921 维的原始向量。因为词典中的很多词在网页中不出现,该网页向量的很多维特征值为 0,即是说该向量极度稀疏。而且原始特征词中的很多词对分类毫无意义,甚至还会引入分类噪声,降低分类精度。比如“如果”“但是”这些在文章中起结构作用的虚词,不表示实际意义,在每篇文章中出现概率大致相等,对分类来说是“平凡词”,应该从特征集中去掉。于是,在进行分类器训练之前,我们必须进行特征选择,选出那些对分类有帮助的词,从而大大压缩特征空间,为后续的分类节省运算时间和存储空间。

特征选择方法常用的文本特征选择方法有:文档频率(DF)[2]、信息增益(IG)[2]、互信息(MI)[2]、X2 统计量(CHI)[2],期望交叉熵[3],文本证据权[3],优势率[3]等。这些方法的基本思想都是对每一个特征(在这里是中文词),计算某种统计度量值,然后设定一个阈值 T,把度量值小于 T 的那些特征过滤掉,剩下的即认为是有效特征。除了介绍 IG,MI,期望交叉熵等经典的特征选择方法之外,这里还介绍了一种改造的优势率方法和一种新的类别区分词的选择方法。

对于特征词 t,各种选择标准的含义如下:

1)文档频数(Document Frequency),即是特征 t 在文本集中出现的文档数。

2)信息增益(Information Gain):

$$IG(t)=-\sum_{i=1}^{m}P(C_i)\log P(C_i)+P(t)\sum_{i=1}^{m}P(C_i|t)\log P(C_i|t)+P(\bar{t})\sum_{i=1}^{m}P(C_i|\bar{t})\log P(C_i|\bar{t})$$

3)互信息(Mutual Information):

$$MI(t)=\sum_{i=1}^{m}P(C_i)\log\frac{P(t|C_i)}{P(t)}$$

$$X^2(t,C_i)=\frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

4)CHI:

$$X^2_{ang}=\sum_{i=1}^{m}P(C_i)X^2(t,C_i)$$

其中 A 是特征 t 和第 i 类文档共同出现的次数,B 是特征 t 出现而第 i 类文档不出现的次数,C 是第 i 类文档出现而特征 t 不出现的次数,D 是第 i 类文档和特征 t 都不出现的次数。

5)期望交叉熵(Expected Cross Entrophy):

$$CE(\mathrm{t}) = P(t)\sum_{i=1}^{m} P(C_i \mid t)\log\frac{P(C_i \mid t)}{P(C_i)}$$

6)文本证据权(Weight of Evidence for Text):

$$WET(\mathrm{t}) = P(t)\sum_{i=1}^{m} P(C_i) \mid \log\frac{P(C_i \mid t)(1-P(C_i))}{P(C_i)(1-P(C_i \mid t))} \mid$$

在以上各式中,$P(C_i)$表示第 i 类文档在文档集合中出现的概率,P(t)表示词 t 出现的概率,$P(\bar{t}) = 1 - P(t)$表示词 t 不出现的概率,$P(C_i \mid t)$表示在出现词 t 的情况下,文档属于第 i 类的概率。$P(C_i \mid \bar{t})$表示词 t 不出现时,文档属于第 i 类的概率。

7)优势率(Odds Ratio)原本用于二元分类器,定义如下:

$$OR(\mathrm{t}) = \log\frac{P(t \mid C_{pos})(1-P(t \mid C_{neg}))}{(1-P(t \mid C_{pos}))P(t \mid C_{neg})}$$

其中:C_{pos}表示正例集的情况,C_{neg}表示负例集的情况。为了适用于多类别的情况,我们提出一种多类别优势率(Multi-Class Odds Ratio)的变体形式如下:

$$MC-OR(\mathrm{t}) = \sum_{i=1}^{m} P(C_i) \times \mid OR(t, C_i) \mid = \sum_{i=1}^{m} P(C_i) \mid \log\frac{P(t \mid C_i)(1-P(t \mid C_{else}))}{P(t \mid C_{else})(1-P(t \mid C_i))} \mid$$

其中,C_{else}表示除第 i 类外的所有类别,即把当前的第 i 类当作正例集,而把所有其他类别合起来作为负例集,从而有$P(\mathrm{t} \mid C_{else}) = \frac{P(t)-P(\mathrm{t} \mid C_i)}{1-P(C_i)}$

8)类别区分词(Category- Discriminating Word)

上述这些选择方法有一个共同的特点:并不按类别计算统计值,选出的是那些全局意义上的“强类别意义”的词,这些词可能有着多类的指示意义。对于不兼类的文本分类问题来说,选用这些词作为分类特征,将使得某些文本向量位于两类的分界线附近,自动分类极易发生错误。于是我们可

以发现这样一种现象,有些词的单类类别意义非常明显,比如“军舰”,“软着陆”,“阿拉法特”,“景泰蓝”等等,它们几乎就只出现在某一类文档之中。比如我们如果要把所有文档分为国际,环保,经济,军事,科教,生活,时政,文娱这八大类,那么“军舰”在文章中的出现就使我们有理由猜测该文章属于军事类,同理,出现“软着陆”的文档极有可能属于经济类。这些词有着极强的类别指示意义,类别区分性相当好,我们称之为“类别区分词”。我们猜测,如果根据词出现的统计信息,选出对应每类的“类别区分词”作为分类的特征表示,那么有可能在大大缩减特征空间的同时,选出那些最具类别指示意义因而也最利于分类的特征。

因此,我们设计“类别区分词”的选取方法如下:

首先,定义词 t1 的类间概率分布如下:

$$\text{Distribute(t)} = (P(C_1 \mid t_1), P(C_2 \mid t_1), ...P(C_n \mid t_1))$$

其中, $P(C_1 \mid t_1) = \dfrac{P(t_1 \mid C_i)P(C_i)}{P(t_1)}$ 为 Bayes 后验概率, $P(t_1) = \sum_{i=1}^{m} P(C_i)P(t_1 \mid C_i)$

而, $P(t_1 \mid C_i) = \dfrac{1 + \sum_{k=1}^{d_i} tf(t_{lk})}{|V| + \sum_{j=1}^{|V|}\sum_{k=1}^{d_i} tf(t_{lk})}$ 表示词 $tf(t_{lk})$ 在 C_i 类的第 k 篇文档中出现的次数。

|V|为总词数, d_i 表示 C_i 类的总文档数。

其次,定义区分词挑选标准 CDW(t)= Max1- Max2,其中 Max1 为 $P(C_1 \mid t_1)$, i= 1,2… ,m 中的最大值,Max2 为次大值。

最后,设置一个阈值 T,T 为 0 到 1 之间的数,CDW(t)值大于 T 的那些词作为类别区分词被挑选出来。

这种做法的直观意义在于,用词与类别之间的后验概率来衡量该词的类别指示意义。而用后验概率最大值和次大值之间的差距来衡量该词的类间区分性。最大类别后验概率越大,与其余类别后验概率之间的差越大,那么该词关于某一类的指示意义就越强。T 的取值一般在 0.3 以上(T 的取值与训练样本集有关),T 越大,选出词的类别区分性越强,但这样的词也越少。由于训练样本集的数量和覆盖广度有限,这种方法选出的类别区分词不会很多。要求的类别区分性越强,这样的词总数越少,但特征太少,表征模式的能力就会大大下降。因此,我们需要在类别区分性和结果词的数量上做一个折衷。

4　文本分类的文本相似度方法和 Naïve Bayes 方法

1)文本相似度法

文本相似度方法其实是一种基于样本相似度的质心分类法。根据待分类的测试样本 di 和各类类中心向量的余弦相似度,把该测试样本判分为相似度最大的那一类。即有:

$$C = \max_{j} Cos(d_i, V_j) = \frac{d_i \times V_j}{|d_i||V_j|} = \frac{\sum_{l=1}^{n} w(t_{il}) w(t_{jl})}{\sum_{l=1}^{n} w(t_{il})^2 \sum_{l=1}^{n} w(t_{jl})^2}$$

2)Naïve Bayes 方法

Naïve Bayes 分类器的一个基本前提是各特征之间的独立性假设,即假定文本中各个特征项属于特定类别的概率相互独立。分类器通过计算待分类样本属于各类的后验概率,把该待分类样本判分为后验概率最大的那一类。Naïve Bayes 分类器的判分准则:

$$C = \max_j P(d \mid C_i)P(C_i) = \max_j P(C_i)\prod_k P(t_k \mid C_i)^{N(t_k,d)}$$

$$P(\mathrm{t_k} \mid C_i) = \frac{1+\sum_{k=1}^{d_i} tf(t_{lk})}{|V|+\sum_{j=1}^{|V|}\sum_{k=1}^{d_i} tf(t_{lk})}$$

其中 , $tf(t_{lk})$ 表示词 $\mathrm{t_i}$ 在 C_i 类的第 1 篇文档中出现的次数，|V|为总词数, d_i 表示 C_i 类的总文档数。 $N(t_k,d)$ 表示词 t_k 在文档 d 中出现的次数。

二、决策树分类算法

所谓决策树，就是一种树结构，这种结构是根据策略选择而构建起来的。在策略选择中涉及到对象属性与对象值之间的映射关系，而这种映射关系由决策树代表，因此从本质上看决策树就是一个预测模型。整个决策树的起始点是决策树的最高级节点，也就是决策树的根节点。某个类别是由树中的叶节点代表，每一种分类结果就是结构树上的每一根分支。决策树常用的分类算法有 ID3 算法、C4.5 算法。其中 ID3 算法是一个启发式算法，C4.5 算法是对 ID3 算法的改进算法，使用信息增益率来选择属性。

三、K-最近邻法

K-Nearest Neighbor 就是 K - 最近邻分类算法，又称 KNN 法。这种算法是一个历史时间久，基于实例的文本分类应用方法，理论上比较成熟的方法，最初由卡瓦和哈特于 1968 年提出的。K - 最近邻分类算法的思路简单直观：对于给定的一个样本，首先找到该样本在特征空间中的 k 个最邻近的样本属于某一个类别，则该样本也属于这个类别。因为该方法中所选择的邻居都已经进行过正确的分类，所以在确定样本类别时，仅仅

根据待定样本最邻近的一个或者几个样本的类别就可以决定待分样本所属的类别。 KNN 方法就是所谓“近朱者赤、近墨者黑”的“懒惰”的学习方法，主要依据周围邻近的有限的样本类别判断自己的类别，不需要通过判别类域来确定所属类别.由于 KNN 方法在判定类别时需要将全部的训练集与测试实例进行相似度比较，并根据相似度进行排序，所以所需分类时间比较多。

四、贝叶斯分类算法

贝叶斯分类算法是所有以贝叶斯定理为基础的分类算法的总称,这种算法是利用概率统计知识进行分类，是基于统计学的一类分类算法。贝叶斯分类算法不仅简单而且分类速度快，分类结果准确率比较高。正是因为这些优点，所以贝叶斯算法常被运用到大型数据库中，可以与决策树和神经网络分类算法相媲美。贝叶斯分类器是使用贝叶斯算法构造出的，是一种应用比较广泛的分类器。 根据贝叶斯分类算法的特点，本文采用朴素贝叶斯分类算法实现校园留言板垃圾留言过滤。

3　基于贝叶斯分类算法的留言板垃圾留言过滤

贝叶斯分类算法是基于统计学的一种分类算法，这类算法都是以贝叶斯定理为依据的分类算法，是一类利用概率统计知识进行分类的算法,是以严谨的数学理论作为支撑。 贝叶斯定理是起源于英国学者贝叶斯创建的贝叶斯统计理论。贝叶斯算法是由传统的概率学理论中脱颖而出的，虽然它专门用于解决统计学中那些具有不确定性的问题，从根本上仍以概率理论为理论基石。

3.1 贝叶斯分类算法

3.1.3 朴素贝叶斯分类算法的应用

根据朴素贝叶斯分类算法原理，可以将分类过程总结为以下步骤：

第一步，为朴素贝叶斯分类做准备工作，主要是根据具体情况确定特征属性，并对每个特征属性进行类别的划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。由人工输入所有待分类数据，输出特征属性和训练样本。

第二步，计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。

第三步，使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。

在实际应用中，把留言看成一个分类（合法留言和垃圾留言）问题。首先收集大量合法留言和垃圾留言当作样本，然后使用贝叶斯分类器对收集到的样本进行有指导的学习，最后使用训练好的贝叶斯分类器对新写入的留言进行分类。通过对留言样本的训练和学习，分类器可以自动获取垃圾留言的特征，并根据垃圾留言特征的变化，准确地对垃圾留言进行过滤。

1　朴素贝叶斯文本分类器

朴素贝叶斯分类器假设特征对于给定类的影响独立于其它特征,即特征独立性假设.对文本分类来说，它假设各个单词W_i和W_j之间两两独立，其原理见图 1 .

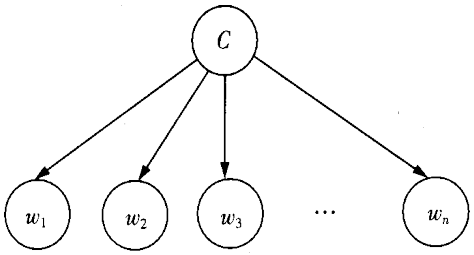


图 1 朴素贝叶斯分类器

设训练样本集分为k类,记为$C=\{C_1,C_2,\ldots,C_k\}$,则每个类C_i的先验概率为$P(C_i),i=1,2,\ldots,k$,其值为C_i类的样本数除以训练集总样本数 n.对于新样本 d,其属于C_i类的条件概率是$P(d\mid C_i)$.根据贝叶斯定理,C_i类的后验概率为$P(C_i\mid d)$:

$$P(C_i\mid d)=\frac{P(d\mid C_i)P(C_i)}{P(d)} \tag{1}$$

$P(d)$对于所有类均为常数,可以忽略,则式(1)简化为

$$P(C_i\mid d)\propto P(d\mid C_i)P(C_i) \tag{2}$$

为避免$P(C_i)$等于 0 ,采用拉普阿斯概率估计:

$$P(C_i)=\frac{1+|D_{Ci}|}{|C|+|D_C|} \tag{3}$$

式中:$|C|$为训练集中类的数目,$|D_{Ci}|$为训练集中属于类C_i的文档数,$|D_C|$为训练集包含的总文档数

在特殊情况下,训练样本集中各类样本数相等,此时类的先验概率相等,式(2)可以简化:

$$P\left(C_i\mid d\right)\propto P(d\mid C_i) \tag{4}$$

朴素贝叶斯分类器将未知样本归于类的依据,如下:

$$P(C_i\mid d)=\arg\max\left\{P(C_j\mid d)P(C_j)\right\},j=1,2,\ldots,k \tag{5}$$

文档d由其包含的特征词表示，即$d=\left(w_1,\ldots,w_j,\ldots,w_m\right)$，m是d的特征词个数d，wj是第j个特征词，由特征独立性假设，则得

$$P(d\mid C_i)=P((w_2,w_2,\ldots,w_m)\mid C_i)=\prod_{j=1}^{m}P(w_j\mid C_i) \tag{6}$$

式中：$P(w_j\mid C_i)$表示分类器预测单词w_j在类C_i的文档中发生的概率.

因此式(2)可转换为

$$P(C_i\mid d)\propto P(C_i)\prod_{j=1}^{|d|}P(w_j\mid C_i) \tag{7}$$

为避免式(7)中$P(w_j\mid C_i)$等于0,可以采用拉普拉斯概率估计.

有两种方法计算$P(w_j\mid C_i)$，即文档型计算公式和词频型计算公式.

文档型:不考虑单词在文档中的出现频次,仅考虑单词在文档中是否出现,0表示未出现,1表示出现,依式(8)计算:

$$P(w_j\mid C_i)=\frac{1+N(doc(w_j)\mid C_i)}{2+|D_c|} \tag{8}$$

式中：$N(doc(w_j)\mid C_i)$为C_i类文本中出现特征的文本数.

2)词频型:考虑单词在文档中出现的频次,依式(9)计算

$$P(w_j\mid C_i)=\frac{1+TF(w_j,C_i)}{|V|+\sum_{k=1}^{|V|}TF(w_k,C_i)} \tag{9}$$

式中：$|V|$表示特征词表中总单词数，$TF(w_j,C_i)$表示单词w_j在类C_i的所有文档中出现的频次之和.本文研究的朴素贝叶斯分类系统公式(7)进行了改进:

$$P(C_i\mid d)\propto P(C_i)\prod_{j=1}^{|d|}P(w_j\mid C_i)^{f(w_j)} \tag{10}$$

式中：$f(w_j)$为单词的评估函数.这样，$f(w_j)$越小，单词w_j在朴素贝叶斯分类器中的作用越小，当$f(w_j)$为 0 时，$P(w_j|C_i)$实际上就不作用了，即$P(w_j|C_i)^{f(w_j)}$等于 1

5　实验结果与结论

为了比较上述八种特征选择方法,我们在两个数据集上作了测试。数据集 1 包含八类共 16000 个网页的样本集,来自人民网 2001 年 1 月到 2003 年 1 月的新闻语料,涵盖国际、经济、军事、环保、科教、社会时政、生活、文娱八大类。各类样本数相等,均为 2000 个。训练集和测试集取 4:1 的比例。即训练集中有 12800 个样本,而测试集中有 3200 个样本。数据集 2 来自 2003 年 3 月在北京大学举办的“中文网页分类竞赛”中给出的训练网页集,共包含 11 类 13890 个网页,各类网页数从最少的 138 到最多的 2841 个,在各类中的网页数分别为:人文与艺术 496,新闻与媒体 138,商业与经济 1024,娱乐与休闲 1846,政府与政治 368,社会与文化 1353,教育 364,自然科学 2255,社会科学 2160,计算机与因特网 1045,医疗与健康 2841,网页在各类的分布极不均匀。训练集和测试集同样取 4：1 的比例。从分类结果我们可以看出,这种网页集类间分布不均的情况对最终的分类结果有一定影响。

为了综合考虑分类精度和召回率,全面衡量分类系统性能,我们使用宏 F1 值作为评价指标,计算如下:

$$Macro\_F1=\sum_{i=1}^{m}\frac{N_i}{N}\times F1_i=\sum_{i=1}^{m}\frac{N_I}{N}\times\frac{2\times precisian_i\times recall_i}{precisian_i+recall_i}$$

其中, N_i 为第 i 类的测试文档数,N 为测试文档总数。 $precisian_i$ 和 $recall_i$ 分别为第 i 类的正确率和召回率,共有 m 个类别。

对于数据集 1,图 1 和图 2 分别给出了采用文本相似度方法和 Naïve Bayes 分类方法时,用上述 7 种特征选择方法(除了 Mutual Information 方法)选出 500～ 20000 维特征时,得到的相应分类性能。横坐标为特征维数 Features,纵坐标为宏 F1 值。

同样,在数据集 2 上用文本相似度方法和 Naïve Bayes 方法,用 7 种特征选择方法作特征选择的效果如图 3,图 4 所示。

在上述各图中没有出现应用 MI 方法进行特征选择的分类结果,是因为 MI 效果太差,在两个数据集上,用文本相似度方法和 Naïve Bayes 两种方法做分类的情况下,在特征维数低于 20000 维时,得到的宏 F1 值都不超过 60%。Yang Yiming 曾对此给出了解释,她认为这是由于 MI 方法在选择特征时,偏爱那些出现频率低的词[2]。为了重点比较其他 7 种特征选择方法的效果,故在图中不再画出 MI 作特征选择的效果.

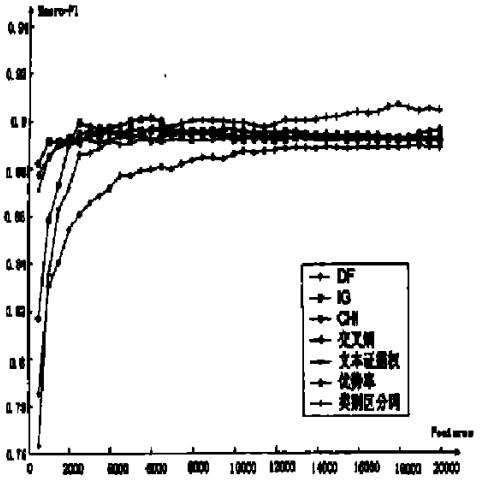


图 1　数据集 1,采用文本相似度方法

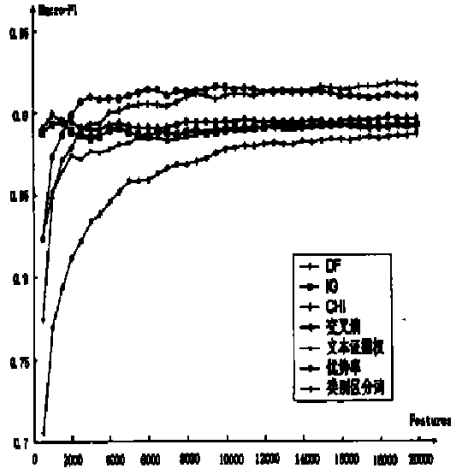


图 2　数据集 1,采用 Naïve Bayes 方法

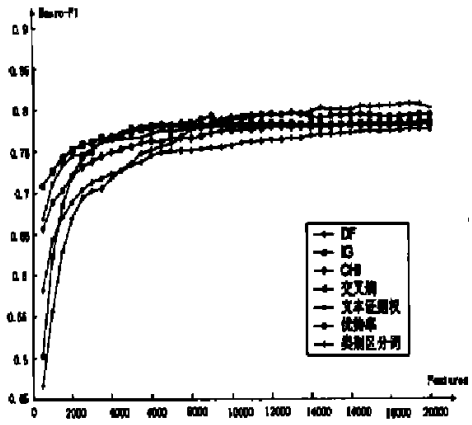


图 3　数据集 2, 采用文本相似度方法

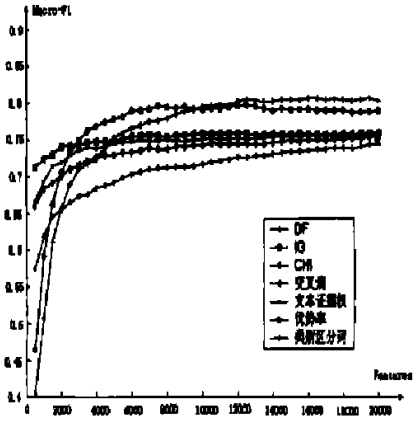


图 4　数据集 2, 采用 Naïve Bayes 方法

由上述诸图,我们可以得出以下结论:

1)改进的多类别优势率和类别区分词,这两种方法效果最好。在不同的数据集和不同的分类方法下,这两种方法的选择效果一直优于其他几种方法。比如在图 1 中,用文本相似度方法对 16000 网页集进行分类训练和测试时,当特征维数大于 6000 维,类别区分词作特征选择的效果最好。

2)从实验结果来看,类别区分词和多类优势率的效果最好, IG 和期望交叉熵其次,文本证据权和 CHI 再次,DF 效果最差。由于只用到了特征词的文档频率信息,因此 DF 的选择效果最差也就不足为奇了。通常我们把 DF 方法作为比较的基准。

3)当数据集是均匀分布时(如图 1 和图 2 所示的 16000 数据集),CHI 作特征选择的效果略优于 IG 和期望交叉熵,而当数据集的类别分布极为不均时(如图 3 和图 4 所示的 13890 数据集),当特征维数低于 10000 维时,IG 和期望交叉熵比 CHI 有着明显的优势,当特征维数高于 10000 维时,CHI,IG,期望交叉熵趋于相同的效果。

4)IG 和期望交叉熵的曲线基本重合,说明这两种方法做特征选择时,有着相似的效果。把 IG 公式重写为如下形式:

$$IG(\mathrm{t}) = P(t)\sum_{i=1}^{m} P(C_i \mid t)\log\frac{P(C_i \mid t)}{P(C_i)} + P(\bar{t})P(C_i \mid \bar{t})\log\frac{P(C_i \mid \bar{t})}{P(C_i)}$$

第一部分就是期望交叉熵,不同的是 IG 还考虑了特征不出现情况下的信息贡献,而从本次实验结果来看,这部分的贡献很小。

5)在图 1 和图 2,图 3 和图 4 之间作比较,我们发现当采用文本相似度方法时,各种特征选择方法的效果相差不大,而当采用 Naïve Bayes 方法时,多类别优势率和类别区分词的方法比其他方法有着明显的优势。如图 4 所示,当用 Naïve Bayes 分类方法对各类分布严重不均的 13890 样本集作训练和测试时,当特征维数大于 10000 以后,用类别区分词作特征选择得到的宏 F1 值比用 IG 作特征选择得到的宏 F1 值高出 5%左右。

6)类别分布不同的数据集,在采用相同分类机制和相同特征选择方法进行训练和测试时,

有着不同的分类结果。比较图 1 和图 3,图 2 和图 4,我们发现各类样本均匀分布的 16000 样本集最高能达到 90%左右的宏 F1 值,而各类样本分布极为不均的 13890 样本集只能达到最高 80%左右的宏 F1 值。这说明样本集的选取对文本分类的绝对结果有着相当大的影响,但同时我们也注意到,一些相对的结果(如上述结论 1- 5)在不同样本集的测试结果中仍成立。

3.2 问题二的建立与求解

3.2.1 问题分析

1.在问题一的基础上再将附件一中每一辆车的不良驾驶行为（主要包括疲劳驾驶、怠速预热、超长怠速、熄火滑行、超速、急变道）都进行相应数据的挖掘，并用层次分析法和聚类算法进行规律的提取和寻找，然后建立科学的数学模型，再进行模型的评估和检验。

2.将上一步的所有模型进行整合，给出一个科学的评价准则，并基础上分析相应不良驾驶行为的影响因素。

3.2.2 问题求解

在向量空间模型中文档被形式化为 n 维空间中的向量，空间的一维是倒排表中的一个元素，形式如下:

$$D=\langle W_{term1},W_{term2},W_{term3},\ldots,W_{termn}\rangle$$

该向量中每一维的值表示该词语在此文档中的权重，用以刻画该词语在描述此文档内容时所起作用的重要程度。词语权重计算唯一的准则就是要最大限度的区分不同文档。所以，针对词语权重的计算,需要考虑 3 个因素

[因素 1] 词语频率 tf:该词语在此文档中出现的频率。

[因素 2] 词语倒排文档频率 idf:该词语在文档集合中分布情况的量化，常用的计算方法是 $\log(N/n_k+0.01)$;其中 N 为文档集合中的文档数目； n_k ：出现过该词语的文档数目。

[因素 3] 归一化因子:对各分量进行标准化。根据上述三个因素，可以得出公式(1) ：

$$W_{ik}=\frac{tf_{ik}\log(N/n_k+0.01)}{\sqrt{\sum_{k=1}^{t}(tf_{ik})^2\left[\log(N/n_k+0.01)\right]^2}} \qquad (1)$$

公式(1)的提出是基于这样一种假设:对区别文档最有意义的词语应该是那些在文档中出现频率足够高，但在整个文档集合的其他文档中出现频率足够少的词语。所以，向量空间模型的基础是词语的出现频率和出现文档频率。

为了说明问题，我们暂且不考虑停用词表。由于在每个文档中助词“的”词频总是最高，所以由 tf.idf 计算所得“的”单字词权重会大于一些实词的权重，换句话讲，“的”单字词在表达文档内容时所传递的信息量要大于一些实词所起的作用。显然，这个结果是不合适的。下面的简单例子(例 1)可以来说明 tf.idf 的优缺点及其原因。设有 2 个文档 text1 和 text2 ，其中的倒排表中仅有 4 个元素。词语在文档中的出现频率如表 1 所示:

表 1　文档词语频率表

文档	词　语			
	term1	term2	term3	term4
text1	20	18	10	0
text2	20	5	0	1

此例在本文中仅用于说明词语分布比例的可计算性和可自动获取的特点。如果我们承认并遵循公式(1)的假设，虽然不能估计文档中各个词语的权重值，但仅跟据词语频率和出现文档频率这一简单信息，并以最大限度的区分文档为最终目标，足以确定词语重要性的排序次序，如表 2 所示

表 2　主观词语权重比较表

文档	词　语　权　重　比　较
text1	$W^{**}_{term3} > W_{term2} > W_{term1}$
Text2	$W_{term4} > W_{term2} > W_{term1}$

由 tf.idf 计算所得的文档向量分别为:

表 3　tf.idf 词语权重计算结果表

文档	文　档　的　词　语　权　重			
	term1	term2	term3	term4
text1	0.028484535830604 96	0 .025636082247544 46	0 .999265441464635 3	0
text2	0 .273493146804377	0 .068373286701094 26	0	0.959440770614178 9

由于 tf.idf 是基于经验的计算模型，我们无法精确判断各个词语的 tf.idf 计算结果是否准确，但我们可以通过文档中各个词语的权重值的比较，通过获取各词语在表示文档时所起作用的重要性序列来考察 tf.idf 计算结果在表示文档时的准确性。tf.idf 计算所得词语权重比较如表 4 所示

文档	词　语　权　重　比　较
text1	$W_{term3} > W_{term1} > W_{term2}$
Text2	$W_{term4} > W_{term1} > W_{term2}$

tf .idf 计算结果分析:

虽然例 1 的数据过于简单，但足以说明问题的所在及其出现原因。对比表 1 和表 4 ，我们可以清楚地了解 tf.idf 词语权重计算公式的利弊。

1 .由于 term3 和 term4 分别在一个且仅在一个文档中出现，尽管出现频率不高，但仍分别在各自文档中被定为最能区别文档的词语。这是 tf.idf 的优点。

2 .term1 在文档 text1 和 text2 中出现频率均是最高，且出现频率相等，即可以认为 term1 对区分二文档毫无意义，本应收入停用词词典中。但在两个文档中W_{term1}均大于W_{term2} 。这正是 tf.idf 的缺点。可见，虽然 tf 在添加 idf 后，为获得文档最大区分能力，对词语在文档集合中分布情况有所考虑，使向量空间模型的文档表示准确度得到很大提高，并在应用问题中得到体现。但由于没有考虑分布的比例情况，才会造成有$W_{term1} > W_{term2}$的不足。所以，为了获得一种最大限度区分不同文档的能力，我们还需考虑另外一个因素:

[因素 4] 词语信息增益:词语在区分文档时所能提供的信息量。

如何获得词语信息增益和添加[因素 4] 后文本表示的提高效果如何的问题，正是本文关注和需要解决的问题。

信息增益的引入

我们认为区分文档的问题可以被形式化为一个分类问题,进而可以把词语在文档中权重计算问题转化为词语在以一个文档为一类的文本分类中的权重计算问题。为了把握词语在各　文档中分布比例对权重计算的影响，我们引入信息论中信息增益的方法来实现这一目标。

在这里把训练数据，即文档集合看作一个符合某种概率分布的信息源，依靠训练数据集合的信息熵和文档中词语的条件熵之间信息量的增益关系确定该词语在分本分类中所能提供的信息量，即词语在分类中的重要程度，并把这种重要程度定义为该词语在文本分类中的权重。见公式(2)：

$$IG_{term}(i) = H(D) - H(D \mid term(i)) \tag{2}$$

公式(2)中的 H(D)为：

$$H(D) = -\sum_{d \in D} P(d) \times \log_2 P(d) \tag{3}$$

其中，

$$P(d) = \frac{|wordser(d)|}{\sum_i |wordser(d_i)|} \tag{4}$$

说明：$wordser(d)$ 表示文档 d 中词语集合的个数。

这样，词语权重计算公式将被改写为公式(4)的形式，公式(4)正是本文提出的文档表示改进方法 tf.idf.IG

$$W_{ik} = \frac{tf_{ik} \times \log(N / n_k + 0.01) \times IG_k}{\sqrt{\sum_{k=1}^{t} \left(tf_{ik} \times \log(N / n_k + 0.01) \times IG_k\right)^2}} \tag{5}$$

例 1 中各个词语对分类问题的信息增益结果、tf.idf .IG 的词语权重计算结果及其排序结果分别在表 5、表 6 和表 7 中给出：

表 5　信息增益计算结果表

$H(D)$	1.0		
$H(D \mid a1)$	1.0	$IG(a1)$	0
$H(D \mid a2)$	0.7553754125614287	$IG(a2)$	0 .2446245874385713
$H(D \mid a3)$	0	$IG(a3)$	1.0
$H(D \mid a4)$	0	$IG(a4)$	1.0

表 6　tf .idf .IG 词语权重计算结果表

文档	文 档 的 词 语 权 重			
	term1	term2	term3	term4
text1	0	0 .00671003078450019 904668697364352959	0 .99997748749002897 722397052184287	0
text2	0	0.0186361569499317079 141342907864603	0	0 .99982633174673766 045472557886748

表 7　tf.idf.IG 词语权重比较表

文档	词 语 权 重 比 较
Text1	Wterm3 > Wterm2 > Wterm1
Text2	Wterm4 > Wterm2 > Wterm1

由表 7 可见, tf.idf.IG 计算所得词语权重的次序与表 2(目标次序)是一致的。可见, 在添加信息增益因子后, tf.idf.IG 不仅兼顾了词语在文档集合中的分布情况 , 而且还考虑了词语在文档集合中的分布比例情况, 使文档通过改进的文档表示方法 tf.idf.IG 计算所得词语权重得以更准确地表现文档内容。

3.3 问题三的分析与求解

4.模型的综合评价

模型的优点:

(1)本文在正确、清楚地分析题意地基础上，建立了对应问题的各个模型。

(2)所有模型都有效的考虑了全局优化的问题，从而同时得到多个未知参数的最优组合解,并且通过软件实现的参数解精度很高，因此得到的结果可信度较高。

(3)不仅如此，本文建立的模型能结合实际情况对问题进行求解，实用性强，具有很好的推广性。

模型的缺点:

(1)模型的不可控因素太多。

(2)无法精确地考虑到客观因素对结论的影响，结果是在较理想的状况下得到的。

(3)本模型因处理数据有限，得到的结果可能与实际值有偏差。

5.参考文献

[1] 中国互联网络信息中心. 第 37 次中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201601/t20160122_53271.htm

[2] 中国互联网络信息中心. 2014 年中国青少年上网行为研究报告 [EB/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg.htm

[3] 周念. 面向用户定制的文本过滤技术研究及应用 [D].北京: 北京邮电大学,2014.

[4] 吴至真. 基于关键词匹配的网页文本过滤算法的研究及实现[D].贵阳: 贵州大学, 2009.

[5] 阮彤,冯东雷,李京. 基于贝叶斯网络的信息过滤模型研究[J]. 计算机研究与发展, 2002(12): 1564 — 1571.

[6] 曹海.基于文本内容分析的过滤技术研究[J]. 四川大学学报:自然科学版,2006, 43(6):1248-1252.

[7] 吕滨，雷国华，于燕飞，杨泽雪，王亚东.基于语义分析的网络不良信息过滤系统研究
[J]. 计算机应用与软件, 2010, 27(2):283-28.
[8] 史忠植.知识发现[M].北京:清华大学出版社,2002.
[9] Yang Yiming,Pederson JO.AComparative Study onFeature Selection inTextCategorization[A].Proceedings of the 14th International Conference on Machine learning[C]. Nashville:Morgan Kaufmann, 1997: 412- 420.
[10] Mlademnic,D., Grobelnik,M. Feature Selection for unbalanced class distribution and Naïve Bayees[A]. Pro-ceedings of the Sixteenth International Conference on Machine Learning[C]. Bled: Morgan Kaufmann, 1999:258- 267.
[11] 王梦云,曹素青.基于字频向量的中文文本自动分类系统[J].情报学报,2000,19(6): 644- 649.
[12] Y.Yang. Noise reduction in a statistical approach to text categorization[A]. Proceedings of the 18th Ann Int ACMSIGIR Conference on Research and Development in Information Retrieval(SIGIR '95)[C]. Seattle: ACMPress, 1995: 256- 263.
[13] 范焱,郑诚,等.用NaïveBayes方法协调分类Web网页[J].软件学报,2001, 12(9): 1386- 1392.
[14] 刘斌,黄铁军,程军,高文.一种新的基于统计的自动文本分类方法[J].中文信息学报,2002,16(6):18- 24.
[15] 梁久祯,兰东俊,扈.基于先验知识的网页特征压缩与线性分类器设计[A].第十二届全国神经计算学术大会论文集[C].北京:人民邮电出版社,2002,494- 501.
[16] Thorsten Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features[A], In: European Conferrence on Machine Learning(ECML)[C]. Berlin: Springer, 1998,137- 142.