

一种高性能的两类中文文本分类方法

樊兴华^{1),2)} 孙茂松¹⁾

¹⁾(清华大学计算机科学与技术系智能技术与系统国家重点实验室 北京 100084)

²⁾(国家知识产权局 北京 100088)

摘 要 提出了一种高性能的两类中文文本分类方法. 该方法采用两步分类策略: 第 1 步以词性为动词、名词、形容词或副词的词语作为特征, 以改进的互信息公式来选择特征, 以朴素贝叶斯分类器进行分类. 利用文本特征估算文本属于两种类型的测度 X 和 Y , 构造二维文本空间, 将文本映射为二维空间中的一个点, 将分类器看作是在二维空间中寻求一条分割直线. 根据文本点到分割直线的距离将二维空间分为可靠和不可靠两部分, 以此评估第 1 步分类结果, 若第 1 步分类可靠, 做出分类决策; 否则进行第 2 步. 第 2 步将文本看作由词性为动词或名词的词语构成的序列, 以该序列中相邻两个词语构成的二元词语串作为特征, 以改进互信息公式来选择特征, 以朴素贝叶斯分类器进行分类. 在由 12600 篇文本构成的数据集上运行的实验表明, 两步文本分类方法达到了较高的分类性能, 精确率、召回率和 F_1 值分别为 97.19%, 93.94% 和 95.54%.

关键词 文本分类; 文本过滤; 高性能; 中文信息处理

中图法分类号 TP18

A High Performance Two-Class Chinese Text Categorization Method

FAN Xing-Hua^{1),2)} SUN Mao-Song¹⁾

¹⁾(State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084)

²⁾(State Intellectual Property Office of the People's Republic of China, Beijing 100088)

Abstract Text filtering for topic-sensitive information is one of the important applications in text categorization. To effectively filter out the topic-sensitive information from Chinese text collections is a technical challenge. This paper presents a high performance method employing a two-step strategy to classify texts. In the first step, authors regard the words with parts of speech verb, noun, adjective and adverb as candidate features, perform feature selection on them in terms of the improved mutual information formula, and classify the input texts with a naive Bayes classifier. A portion of texts which are currently thought of being unreliable in categorization are identified, forming a fuzzy area between categories. In the second step, authors regard the bigrams of words with parts of speech verb and noun as candidate features, use the same feature selection and classifier to deal with the texts in the fuzzy area. The experiments on a test set consisting of 12600 Chinese texts show that this method achieves a high performance. The precision, recall and F_1 is 97.19%, 93.94% and 95.54% respectively.

Keywords text categorization; text filtering; high performance; Chinese information processing

1 引 言

随着 Internet 的迅速发展, 互联网上的信息极

大丰富, 已使它成为全球最大的分布式信息库. 目前绝大多数信息均表现为文本方式, 如何在浩如烟海而又纷繁芜杂的文本中掌握最有效的信息始终是信息处理的一大目标. 由于分类可以在一定程度上解

决网上信息杂乱的现象,方便用户准确地定位所需的信息和分流信息,因此,文本自动分类已成为一项具有较大实用价值的关键技术,是组织和管理数据的有力手段.文本分类的方法很多^[1],典型的有朴素贝叶斯分类器^[2]、基于向量空间模型的分分类器^[3]、基于实例的分类器^[4]和用支撑向量机建立的分分类器^[5,6]等.

有关国家安全敏感信息的文本过滤^[7]是文本分类技术的重要应用之一.该问题可看作一个两类文本分类问题,可形式化地定义为:假设预定义的文本类型集为 $C = \{c_1, c_2\}$,其中 c_1 表示和国家安全敏感信息相关的文本类型, c_2 表示和国家安全敏感信息不相关的文本类型,要进行分类的文本集为 $D = \{d_1, d_2, \dots, d_n\}$,则该任务就是给文本集 D 中的文档 $d_i (i=1, 2, \dots, n)$ 分配一个类型标记 c_1 或者 c_2 ,然后将标记为 c_1 的文本过滤掉,将标记为 c_2 的文本发送给用户.

本文的研究对象限定为一类和国家安全敏感信息相关的真实的中文文本集.该类语料有如下显著特点:(1)文本集中属于类型 c_2 的文本数目远远大于属于类型 c_1 的文本数目;(2)属于类型 c_1 的文本为一些宣扬对国家安全有害内容的文本,而在属于类型 c_2 的文本中存在许多揭露这些对国家安全有害内容的文本,假定这些文本的类型为 c_2 的一个子类 c_2^1 .它们虽属不同的类型,但使用的词语中有相当部分是相同的;(3)在属于类型 c_2 的文本中还存在着许多文本,虽然它们的内容与那些对国家安全有害的内容完全不同,但它们使用的词语中也有相当部分是相同的,假定这些文本的类型为 c_2 的一个子类 c_2^2 .由于问题的特殊性,对这类语料的分类性能提出了更高的要求:(1)要达到较高的召回率,尽可能多地过滤掉属于类型 c_1 的文本,以保证对国家安全有害内容的文本不被传播;(2)要达到较高的精确率,保证在过滤有害信息的同时,那些属于类型 c_2^1 和 c_2^2 的文本不被错误地过滤掉,避免因噎废食,妨碍用户获取信息和进行信息交流.

通用文本分类方法通常采用的是单步分类策略,即首先从待分类文本中抽取具有分辨能力的特征,然后根据某一种分类算法对文本进行分类.对本文研究的这类语料,因其具有如上所述的特点,通用文本方法效果往往不理想,达不到对这类语料分类的性能要求.本文针对这类语料的分类问题提出了一种采用两步分类策略的高性能的分类方法.其基本思路是:第1步,从待分类文本中抽取特征,然后根据某一分类算法进行分类,并对分类结果进行评

估,如果分类结果可靠,则做出分类判断,如果不可靠,则进行第2步分类处理;第2步,从待分类文本中抽取与第1步所使用特征完全不同的特征进行分类,根据两步的分类结果做出最终判断.

2 基于朴素贝叶斯的两步中文文本分类方法

2.1 两类朴素贝叶斯分类器的改写

给定二值文本向量 $d = (W_1, W_2, \dots, W_D)$, $W_i = 0$ 或者 1 ,如果第 i 个特征出现在文本中, $W_i = 1$,否则 $W_i = 0$.令 $p_{ki} = P(W_k = 1 | c_i)$, $P(\cdot)$ 表示求事件 (\cdot) 发生的概率.两类朴素贝叶斯分类器的判别函数可表示为

$$f(d) = \log \frac{P(c_1 | d)}{P(c_2 | d)} = \log \frac{P(c_1)}{P(c_2)} + \sum_{k=1}^{|D|} \log \frac{1 - p_{k1}}{1 - p_{k2}} + \sum_{k=1}^{|D|} W_k \log \frac{p_{k1}}{1 - p_{k1}} - \sum_{k=1}^{|D|} W_k \log \frac{p_{k2}}{1 - p_{k2}} \quad (1)$$

当 $f(d) \geq 0$ 时,文本 d 属于类型 c_1 ;否则属于类型 c_2 .令

$$Con = \log \frac{P(c_1)}{P(c_2)} + \sum_{k=1}^{|D|} \log \frac{1 - p_{k1}}{1 - p_{k2}} \quad (2)$$

$$X = \sum_{k=1}^{|D|} W_k \log \frac{p_{k1}}{1 - p_{k1}} \quad (3)$$

$$Y = \sum_{k=1}^{|D|} W_k \log \frac{p_{k2}}{1 - p_{k2}} \quad (4)$$

Con 只与所采用的训练样本集有关,不随文本 d 的变化而变化, D 为常数; X 表示根据特征估算出来的文本 d 属于类型 c_1 的测度; Y 表示根据特征估算出来的文本 d 属于类型 c_2 的测度,则式(1)可改写为

$$f(d) = X - Y + Con \quad (5)$$

式(5)表示两类朴素贝叶斯分类器可看作是在由 X 和 Y 构成的二维空间中寻求一条分割直线 $f(d) = 0$.这样,利用式(3)和(4),可将文本表示为二维空间中的一个点 (x, y) ,该点到分割直线 $f(d) = 0$ 的距离 $Dist$ 为

$$Dist = \frac{1}{\sqrt{2}}(x - y + Con) \quad (6)$$

如图1所示,当 $Dist \geq 0$ 时,表示文本 d 属于类型 c_1 ;当 $dist < 0$ 时,表示文本 d 属于类型 c_2 .

我们将公式(1)改写为公式(5),再演变为公式(6)的目的是:(1)利用公式(6)可以在由 X 和 Y 构成的二维空间中方便地考察、分析文本分类错误,探讨在给定分类方法和文本特征集的条件下,距离 $Dist$ 与分类错误的关系;(2)利用公式(6)可以根据

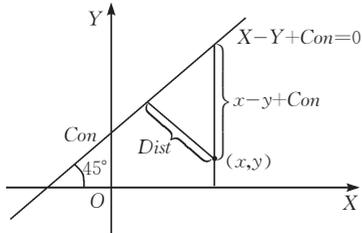
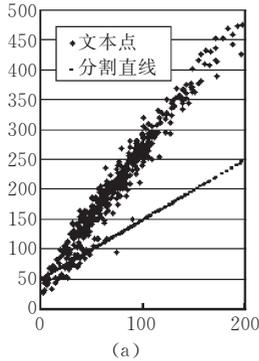
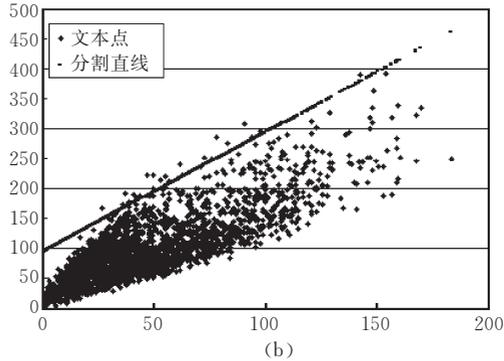


图 1 文本点到分割直线的距离计算

距离 $Dist$ 的大小方便地评估分类的可靠程度, 确定第 1 步分类结果中的不可靠部分, 为后文的两步分类策略的实现做准备。



(a)



(b)

图 2 在由 X 和 Y 构成的二维空间中, 文本点的分布情况

假设: 文本分类器的性能与由公式(6)计算得到的文本到分割直线的距离 $Dist$ 有关, 大多数的错误发生在一个距离分割直线很近的狭窄区域内。也就是说, 如果将待分类文本集中到分割直线距离很近的文本去掉, 那么分类器在由剩余文本构成的新文本集上的测试性能将会提高。

根据假设, 可将由 X 和 Y 构成的二维平面分成可靠和不可靠两个区域。根据式(7)进行分类判别。

$$\begin{cases} Dist_2 \leq Dist \leq Dist_1, & \text{对文本 } d \text{ 的任何分类决策} \\ & \text{都是不可靠的} \\ Dist > Dist_1, & \text{文本 } d \text{ 属于类型 } c_1 \text{ 且分类结果可靠} \\ Dist < Dist_2, & \text{文本 } d \text{ 属于类型 } c_2 \text{ 且分类结果可靠} \end{cases} \quad (7)$$

式(7)中, $Dist_1$ 和 $Dist_2$ 是由实验确定的两个分界常数, $Dist_1$ 为正实数, $Dist_2$ 为负实数。

将文本错误分类的一个原因是, 相对这些文本而言, 所使用的特征不显著。根据假设, 可将分类结果分成分类可靠和不可靠两部分, 这就给我们提供了一种改进分类器性能的机会, 即我们可以对分类不可靠的文本选用更加显著的特征进行二次分类。

2.3 基于朴素贝叶斯的两步文本分类方法

本方法的基本过程为:

1. 以词性为动词、名词、形容词或副词的词语作为特征, 以改进互信息公式(8)选择特征, 以朴素贝叶斯分类器进

2.2 错误分类的文本观察

以第 3 节实验中所使用的语料为样本, 以 X 为横坐标、 Y 为纵坐标, 我们统计出了文本点在二维空间中的分布情况, 如图 2 所示。图 2(a) 对应类型 c_1 的文本分布, 图 2(b) 对应类型 c_2 的文本分布。从图中可以看出, 在二维空间中两类文本以条带形状分布在分割直线的两边; 被错误分类的文本(即图 2(a)中位于分割直线上方的文本点, 或者图 2(b)中位于分割直线下方的文本点)到分割直线的距离很近。根据观察可作如下假设。

行分类。然后根据式(7)评估分类结果, 若分类可靠, 做出分类决策; 否则进行第 2 步。

2. 将文本看作由词性为动词或名词的词语构成的序列, 以该序列中相邻两个词性为动词或名词构成的二元词语串作为特征, 以改进互信息公式(8)选择特征, 以朴素贝叶斯分类器进行分类。

两步分类方法中的几个关键问题如下:

(1) 特征的确定

第 1 步选用的特征应该保证: 具有较强的覆盖能力, 能涵盖所有文本; 具有较强的分辨能力, 能对大多数文本进行可靠的分类, 使系统具有较高效率。第 2 步选用的特征应该保证: 对第 1 步分类结果中不可靠的那些文本具有较强的分辨能力, 使系统具有较高性能。在中文文本分类中, 一般可以选择字、词或者词组作为特征。根据实验结果, 通常认为选取词作为特征要优于字和词组。因此, 我们选用词性为动词、名词、形容词或副词的词语为第 1 步分类的特征。通过对错误文本的观察, 将文本看成由词性为名词或动词构成的词语序列, 选择两个相邻的二元词语串作为第 2 步分类的特征。从中文文本中抽取上述两类特征, 需要进行分词。文中选用了由清华大学开发的汉语分词系统 CSegTag3.0 进行分词。

(2) 特征的选择

由于表示文本的特征向量空间维数相当大。例如在我们后面运行的实验中, 未经特征选择的

词语数目为 69440, 二元词语串数目为 1222482. 为了提高效率, 避免分类器的过度适合问题, 需要进行特征选择. 特征选择的方法很多, 如采用信息增益^[8] (information gain) 和互信息^[9] (mutual information) 等信息理论函数进行特征选择. 文中采用了改进的互信息公式(8)^[10]进行特征选择, 并在实验中比较了它与采用常规互信息公式(9)进行特征选择的性能差异.

$$MI_1(t_k, c) = \sum_{i=1}^n P(t_k, c_i) \log \frac{P(t_k, c_i)}{P(t_k)P(c_i)} \quad (8)$$

$$MI_2(t_k, c) = \sum_{i=1}^n \log \frac{P(t_k, c_i)}{P(t_k)P(c_i)} \quad (9)$$

上面两式中, t_k 表示第 k 个特征, 文中为一个词语或者一个二元词语串; c_i 表示文本预定义类型中的第 i 个类型, 文中 i 取值为 1 或者 2, 类型意义定义同引言中的定义; $MI(t_k, c_i)$ 表示特征 t_k 和类型 c_i 之间的互信息; $P(t_k)$ 表示特征 t_k 发生的概率; $P(c_i)$ 表示类型 c_i 发生的概率; $P(t_k, c_i)$ 表示特征 t_k 和类型 c_i 共现的概率.

有关特征选择的另一个问题是特征集规模的确定, 文中通过分类器的性能随特征数目的变化曲线来确定.

(3) 第 1 步分类结果的评估

对分类结果进行评估的关键是选择两个合适的分界常数 $Dist_1$ 和 $Dist_2$, 将二维空间平面分为可靠和不可靠两个区域. 若不可靠区域太大, 则需要进行分类的文本多, 系统效率低; 若不可靠区域太小, 对分类性能的改善将会有限. 文中通过实验进行参数确定.

2.4 两步分类方法的特点

两步分类方法的一个显著特点是能够组合两种不同类型的特征进行分类. 在文本分类中, 人们通常使用词语作为特征表示文本, 而那些具有更好语义品质的词语组合, 因其具有较差的统计品质而导致分类性能相对较差被弃用^[11]. 两步分类方法有效地将两种特征组合起来, 以词语进行第 1 步分类, 对用词语不能可靠分类的那些文本, 采用具有更好语义品质的词语组合进行分类. 这样就利用了词语特征的统计品质和词语组合特征的语义品质.

此外, 两步分类方法和分类器委员会方法在思路上是不同的, 前者是采用同一种分类方法和不同的特征进行决策, 两次分类之间是一种串联关系; 后者是通过不同分类器进行综合决策, 各个分类器之间是一种并联关系.

3 实验

3.1 实验数据集

本文用于实验的数据集收集文本共 12600 篇, 其中宣扬对国家安全有害内容的文本为 1800 篇, 它们构成属于类型 c_1 的文本集; 揭露这种对国家安全有害内容的文本为 3716 篇, 它们构成属于类型 c_2^1 的文本集; 内容与那些对国家安全有害的内容完全不同, 但它们使用的词语中有相当部分是相同的文本为 828 篇, 它们构成属于类型 c_2^2 的文本集; 其它文本为 6256 篇, 它们构成属于类型 c_2^3 的文本集; 文本集 c_2^1, c_2^2 和 c_2^3 共同构成属于类型 c_2 的文本集, 共 10800 篇. 为了模拟现实环境中两类文本出现的实际情况, 属于类型 c_1 和属于类型 c_2 的文本数目比例为 1 : 6. 将属于类型 c_1 和属于类型 c_2 的文本集随机地平均分为四份, 以其中的一份构成测试集, 另外的三份构成训练集, 按四栏进行交叉验证, 以四栏实验的平均值作为最终的性能指标.

3.2 分类性能评估指标

对文本分类器的性能采用如下 4 种指标进行评估.

精确率 (Precision):

$$P = \frac{\text{正确分为某类的文本数}}{\text{测试集中分为该类型的文本总数}} \times 100\%;$$

召回率 (Recall):

$$R = \frac{\text{正确分为某类的文本数}}{\text{测试集中属于该类型的文本总数}} \times 100\%;$$

F_1 测试值:

$$F_1 = \frac{2 \times P \times R}{P + R}.$$

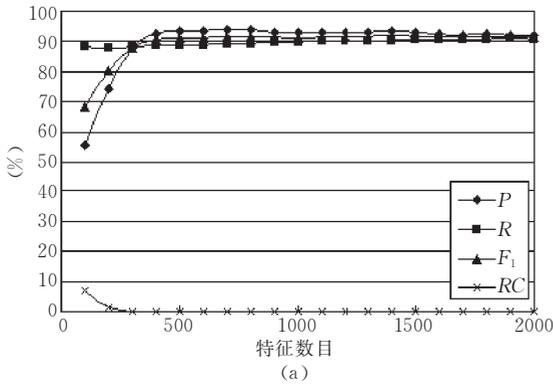
注意到式(3)和式(4)中, 当 W_k 为 0 时, 该特征对 X 和 Y 的值没有贡献, 即文本分类正确与否与待分类文本中存在的有效特征 (也就是 $W_k = 1$ 的特征) 数有关. 这样就可能出现一种情况: 某些待分类文本中根本没有出现任何有效特征. 此时, 分类器作出的任何分类决策实际上没有考虑到文本内容, 是不可靠的. 文中, 我们将这些文本作为一个单独的类型进行标记, 在进行精确率和召回率的计算时, 按训练集中两类文本的分布比例分配到两类文本集中, 即 6/7 的文本算作 c_2 类型, 1/7 的文本算作 c_1 类型. 这类情况的出现反映了所选择特征集对文本的覆盖程度, 为此引入评价指标拒分率. 如果某待分类文本中存在的有效特征数小于 5, 则认为分类器不能对此文本进行分类, 即拒绝分类此文本.

拒分率 (Refuse Category):

$$RC = \frac{\text{属于某类而分类器拒分的文本数}}{\text{测试集中属于该类型的文本总数}} \times 100\%$$

3.3 实验结果

由于测试集中,属于类型 c_1 和属于类型 c_2 的文档比例为 1 : 6,如果将所有文本都标记为 c_2 ,类型 c_2 的分类精度也能达到 85.7%,因此类型 c_2 的分类性能对所选择的分类方法不敏感.为了节约篇幅,在下面的实验中只给出类型 c_1 的分类性能.由分类性能随特征数变化曲线确定特征集规模的标准为:在保证分类性能的同时所使用的特征数应尽可能少,以保证较高的系统性能.因此,通常选择曲线上第 1 个拐点附近的特征数为特征集规模.



实验 1. 特征选择公式的选择和特征集规模的确定.

以词语为特征,以改进的互信息公式(8)和常规的互信息公式(9)分别进行特征选择,采用朴素贝叶斯分类器进行分类,画出了分类性能随特征数目变化的曲线,如图 3 所示,其中图 3(a)对应公式(8),图 3(b)对应公式(9).

由图 3 确定,使用公式(8)时特征集规模为 500 比较合适;使用公式(9)时特征集规模为 4000 比较合适.此时它们的拒分率都为 0,所选特征集能够覆盖所有待分类文本.以选定的特征集规模对两个公式进行定量比较,结果如表 1 所示.

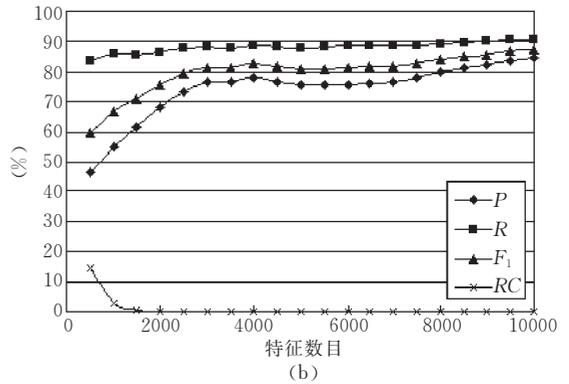


图 3 采用不同特征选择公式,分类器性能随特征数目的变化曲线

表 1 两个特征选择公式的性能比较

使用公式	性能指标		
	精确率 $P(\%)$	召回率 $R(\%)$	$F_1(\%)$
公式(8)	93.35	88.78	91.00
公式(9)	78.04	88.72	82.67

由图 3 和表 1 可以看出,改进的互信息公式(8)比常规互信息公式(9)好,使用它所需特征少,系统效率高;而且精确率有大幅提高,导致总的 F_1 值提高高达 8.33%.

实验 2. 假设的验证和两个分界常数 $Dist_1$ 和 $Dist_2$ 的确定.

为了验证假设,我们引入错误率和区域百分比两个评估指标,定义如下.

错误率(Error Rate):

$$ER = \frac{\text{在某区域内错误分类的文本总数}}{\text{实验中错误分类的文本总数}} \times 100\%$$

区域百分比(Region Per):

$$RP = \frac{\text{在某区域内的文本总数}}{\text{测试集中的文本总数}} \times 100\%$$

当实数 $Dist^* > 0$ 时,将二维文本空间中所有到分割直线距离 $Dist^* > Dist > 0$ 的样本空间定义为区域 A,在此区域内所有的分类错误是将属于类型

c_1 的文本错误的分为类型 c_2 ,这就降低了类型 c_1 的召回率和类型 c_2 的精确率;当实数 $Dist^* < 0$ 时,将二维文本空间中所有到分割直线距离 $0 > Dist > Dist^*$ 的样本空间定义为区域 B,在此区域内所有的分类错误是将属于类型 c_2 的文本错误的分为类型 c_1 ,这就降低类型 c_2 的精确率和类型 c_1 的召回率.

以词语为特征,以改进互信息公式(8)进行特征选择,采用朴素贝叶斯分类器进行分类,画出了 ER 和 RP 随距离 $Dist$ 变化的曲线,如图 4 所示.

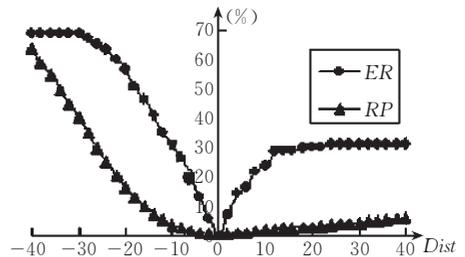


图 4 ER 和 RP 随距离 $Dist$ 的变化曲线

采用和确定特征集规模一样的原则,由图 4 确定分界常数 $Dist_1$ 为 14, $Dist_2$ 为 -28 比较合适.表 2 定量地给出了当 $Dist$ 为 $Dist_1$ 和 $Dist_2$ 时的性能比较.

表 2 给定区域的错误率和区域百分比

区域	错误率(%)	百分比(%)
$14 \geq Dist \geq 0$	28.56	2.08
$0 \geq Dist \geq -28$	66.85	34.79
$14 \geq Dist \geq -28$	95.21	36.87

由图 4 和表 2 可以看出,假设是完全成立的. 95.21%的错误出现的区域内的所有文本仅占文本总数的 36.87%. 其中,66.85%的错误为将属于类型 c_2 的文本错误地分为类型 c_1 , 28.56%的错误为将属于类型 c_1 的文本错误地分为类型 c_2 .

实验 3. 两步文本分类方法的性能及其参数确定.

第 1 步以词语为特征,第 2 步以二元词语串为特征,以改进互信息公式(8)进行特征选择,第 1 步使用的特征集规模为 500,画出了分类器性能随第 2 步特征数目变化的曲线,如图 5 所示.

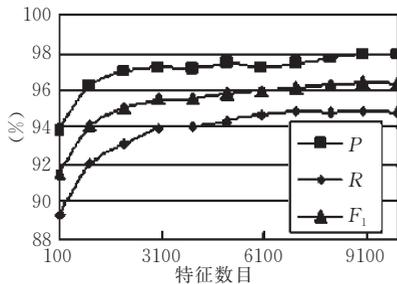


图 5 固定第一步特征规模时,分类器性能随第二步特征数目变化曲线

由图 5 确定,第 2 步特征集规模为 3000 比较合适,此时分类器的精确率、召回率和 F_1 值分别为 97.19%,93.94%和 95.54%.

实验 4. 特征二元词语串的特性分析.

以二元词语串为特征,以改进的互信息公式(8)进行特征选择,采用朴素贝叶斯分类器进行分类,画出了分类性能随特征数目变化的曲线,如图 6 所示. 其中图 6(a)对应性能指标精确率、召回率和 F_1 值;图 6(b)对应类型 c_1 的拒分率 RC_1 和类型 c_2 的拒分率 RC_2 .

由图 6 可以看出,二元词语串是具有较强类型分辨能力的特征,但它也是数据分布很稀疏的一种特征.图 6(a)中,由于它具有较强类型分辨能力,在特征数为 15000 时,精确率为 93.15%,召回率为 94.17%, F_1 为 93.65%,达到了较高的性能.从图 6(b)中可以看出,由于它的数据稀疏性,许多文本中没有有效特征,导致拒分率 RC_1 和 RC_2 较高.在特征数为 15000 时, RC_1 为 0.44%, RC_2 为 6.70%.这说明二元词语串不适合单独作为特征用于分类.

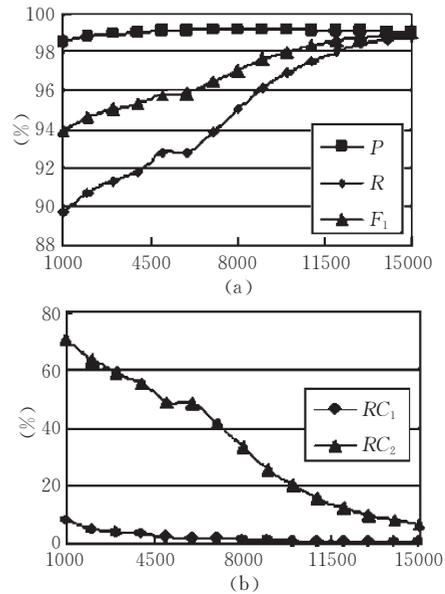


图 6 仅以二元词语串为特征的分类性能随特征数目变化曲线

实验 5. 5 种分类方法的性能比较.

方法 1. 以词语为特征,以常规的互信息公式(9)进行特征选择的单步分类;

方法 2. 以词语为特征,以改进的互信息公式(8)进行特征选择的单步分类;

方法 3. 以二元词语串为特征,以改进的互信息公式(8)进行特征选择的单步分类;

方法 4. 以词语和二元词语串作为特征,以改进的互信息公式(8)进行特征选择的单步分类;

方法 5. 以词语和二元词语串分别作为第 1 步和第 2 步的特征,以改进的互信息公式(8)进行特征选择的两步分类;

实验结果如表 3 所示. 其中,方法 1 和方法 2 的实验结果来源于本节的实验 1;方法 3 的实验结果来源于本节的实验 4;方法 5 的实验结果来源于本节的实验 3. 采用方法 4,画出了分类性能随特征数目变化的曲线,如图 7 所示. 由图 7 确定,混合特征集的规模为 800 比较合适. 此时分类性能如表 3 所示.

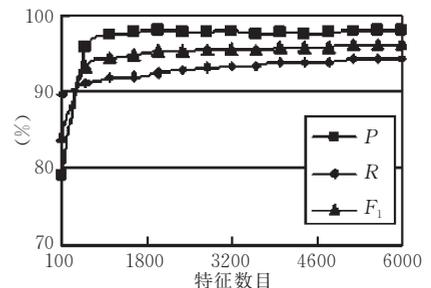


图 7 以词语和二元词语串作为特征时,分类器性能随特征数目变化曲线

表 3 5 种方法的性能比较

使用方法	性能指标			使用的特征数
	精确率 $P(\%)$	召回率 $R(\%)$	$F_1(\%)$	
方法 1	78.04	88.72	82.67	4000
方法 2	93.35	88.78	91.00	500
方法 3	93.15	94.17	93.65	15000
方法 4	95.86	91.11	93.42	800
方法 5	97.19	93.94	95.54	500+3000

由表 3 数据,比较方法 1 与方法 2、方法 3、方法 4 和方法 5 可以看出,特征选择公式(8)优于特征选择公式(9),按照 F_1 值,采用其前者分类器的性能明显大幅度高于后者。

比较方法 2 与方法 3、方法 4 和方法 5 可以看出,二元词语串具有较强的类型分辨能力,以其为特征,按照 F_1 值,分类器的性能高于仅以词语为特征的情况。

比较方法 3 与方法 4 和方法 5 可以看出,以词语和二元词语串为特征优于仅以二元词语串为特征,尽管方法 4 和方法 3 按照 F_1 值分类器的性能差不多,但前者使用的特征数比后者少,系统效率高。

与其它 4 种方法相比,按照 F_1 值,方法 5(即两步分类方法)的性能最好。方法 5 中仅仅在第 2 步分类中使用了二元词语串,一方面利用了二元词语串具有较强类型分辨能力的特性,另一方面也避免了因数据稀疏带来的文本拒分问题且提高了系统效率,因为第 2 步只对第 1 步分类结果评判中不可靠的一小部分文本进行二次分类,从本节的实验 2 可知这一小部分文本只占总文本数的 36.87%。

4 结 论

有关国家安全敏感信息的文本过滤是文本分类的重要应用之一。本文针对当前一项急需解决的现实任务,提出了一种高性能的两类中文文本分类方法。该方法采用两步分类策略:第 1 步以词性为动词、名词、形容词或副词的词语为特征,采用改进的互信息公式进行特征选择,以朴素贝叶斯分类器进行分类。利用文本特征估算文本属于两种类型的测度 X 和 Y ,以此构造二维文本空间,将文本映射为二维空间中的一个点,将分类器看作是在二维空间中寻求一条分割直线。根据文本点到分割直线的距离将二维空间分为可靠和不可靠两部分,以此对第 1 步分类结果进行评估。若第 1 步分类结果可靠,作出分类决策;否则进行第 2 步。第 2 步将文本看作由词性为动词或名词的词语构成的序列,以该序列中

相邻两个词语构成的二元词语串为特征,采用改进的互信息公式进行特征选择,以朴素贝叶斯分类器进行分类。在由 12600 篇文本构成的数据集上运行的实验表明,两步文本分类方法能达到较高的分类性能,精确率、召回率和 F_1 值分别为 97.19%, 93.94% 和 95.54%。由本文的实验还能得到其它如下结论:(1)文本分类器的性能与构造的二维文本空间中文本点到分割直线的距离有关,大多数的分类错误发生在一个距离分割直线很小的狭窄区域内;(2)二元词语串是一种具有较强类型分辨能力的特征,它也是数据分布很稀疏的一种特征,不适合单独作为特征用于分类;(3)改进的互信息公式比常规的互信息公式更能选择到那些具有较强类型分辨能力的特征,有效地改善分类器的性能。

参 考 文 献

- 1 Sebastiani F. . Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1~47
- 2 Lewis D. . Naive bayes at forty: The independence assumption in information retrieval. In: Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 1998, 4~15
- 3 Salton G. . Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading, MA: Addison-Wesley, 1989
- 4 Mitchell T. M. . Machine Learning. New York: McCraw Hill, 1996
- 5 Joachims T. . Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 1998, 137~142
- 6 Yang Y. , Liu X. . A Re-examination of text categorization methods. In: Proceedings of SIGIR'99, the 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999, 42~49
- 7 Fan Xing-Hua. Causality reasoning and text categorization. Postdoctoral Research Report of Tsinghua University, 2004(in Chinese)
(樊兴华. 因果推理和文本分类. 清华大学博士后出站报告, 2004)
- 8 Larkey L. S. . Automatic essay grading using text categorization techniques. In: Proceedings of SIGIR'98, the 21st ACM International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998, 90~95
- 9 Dumais S. T. , Platt J. , Hecherman D. , Sahami M. . Inductive learning algorithms and representation for text categorization. In: Proceedings of CIKM'98, the 7th ACM International Conference on Information and Knowledge Management, Bethesda, MD, 1998, 148~155

- 10 Sahami M., Dumais S., Hecherman D., Horvitz E.. A Bayesian approach to filtering junk E-mail. AAAI, Madison Wisconsin; AAAI Technical Report WS-98-05, 1998, 55~62
- 11 Lewis D. D.. An evaluation of phrasal and clustered representa-

tions on a text categorization task. In: Proceedings of SIGIR'92, the 15th ACM International Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 1992, 37~50



FAN Xing-Hua, born in 1972, Ph.D., associate research fellow. His research interests include artificial intelligence, natural language processing, information retrieval, uncertain reasoning and fault diagnosis to complex system.

SUN Mao-Song, born in 1962, Ph.D., professor. His research interests include artificial intelligence, natural language processing and information retrieval.

Background

This work is a part of the "Research on Automated Categorization of Chinese Texts with High Accuracy", which is mainly supported by the National High Technology Research and Development Program (863 Program) and the National Natural Science Foundation of China. As one of the key tech-

niques in information retrieval and knowledge management, text categorization (TC) has attracted much research interests in the past decade. How to further improve the accuracy of categorization is a big challenge in TC. Targeting at this problem, a two-step approach is studied in the paper.

第一届全国数据库应用技术年会

2006 年 10 月 山东济南

主办单位：中国计算机学会数据库专业委员会

承办单位：山东大学计算机科学与技术学院

协办单位：山东地纬计算机软件有限公司

由中国计算机学会数据库专业委员会主办的第一届中国数据库应用技术年会(CDAT2006)(<http://www.sdu.edu.cn/cdat2006>)将于 2006 年 10 月在济南举行.作为对 NDBC 的强力补充,本次会议将展示全国数据库应用的最新技术和成果,为数据库研究者、开发者和用户提供一个数据库应用技术论坛,探讨数据库应用技术所面临的关键问题和发展方向.

CDAT2006 的议题涉及数据库应用及应用平台的多个方面,届时国内外著名专家将到会作专题报告,同时邀请世界著名的数据库厂商就最新的应用工具和平台技术进行交流,满足国内在该领域日益增长的技术和应用需求.

CDAT2006 诚邀各行业数据库工作者踊跃投稿!

征文范围

会议的主要方向包括(不限于此):

大规模数据库应用	数据库存储技术及实现	数据库备份技术	数据库平台和深度挖掘
数据库容灾技术	XML 数据库实现技术	查询处理技术	事务(日志)管理
分析型数据库系统实现技术	数据挖掘的实用技术	数据仓库的实用技术	中间件和应用服务器技术及应用
多媒体数据库技术及应用	生物信息系统应用		

投稿要求

作者投往本届大会的稿件必须是未发表的技术成果、工作经验,论文应包括题目、摘要、关键词、正文和参考文献.作者信息包括论文题目、作者全名、所属单位、电子邮件、通信地址、电话和传真.稿件以 pdf 格式提交,所有稿件进行统一审理.

经评审录用稿件将在《计算机科学》专刊(本次大会论文集)和《山东大学学报》正刊发表,优秀稿件将推荐到国内一级学报正刊发表.

本次会议网址: <http://www.sdu.edu.cn/cdat2006>

联系信箱: cdat2006@sdu.edu.cn

联系人: 宋婷婷 电话: 0531-88169988 转 6111

传真: 0531-88113508

通信地址: 济南市山大南路 27 号山东大学计算机科学与技术学院(邮编:250100) 洪晓光(信函请注明 cdat2006 字样)

重要日期

征稿截止时间: 2006 年 4 月 25 日

论文录用通知时间: 2006 年 6 月 25 日