新型快速中文文本分类器的设计与实现

陈艳秋,熊耀华
CHEN Yan-qiu,XIONG Yao-hua

东北大学 东软信息技术学院 计算机科学与技术系,辽宁 大连 100623
Department of Computer Science,Neusoft Institute of Information,Dalian,Liaoning 100623,China
E-mail:chenyanqiu@neusoft.edu.cn

CHEN Yan-qiu,XIONG Yao-hua.Design and implementation of new Chinese text classier.Computer Engineering and Applications,2009,45(22):53-55.

Abstract: For improving the efficiency and accuracy of Chinese text categorization,this paper presents a new Chinese text classier,in which a novel feature selection is proposed according to word frequency,mutual information and classificatory information,and after analyzing the hypostasis of the traditional TF-IDF,a weight adjustment method is put forward in which the IDF function is replaced by function used in feature selection.Finally a fast Bayes theory classier is designed.Experiments prove this classier is simple and effective.

Key words: Chinese text categorization;feature selection;feature weighting;classification algorithm

摘 要:为了提高中文文本分类的效率与精度,设计了一种新型的分类器。该分类器采用基于词频、互信息和类别信息的综合评估函数进行选择特征;在特征权重计算上,由于传统 TF-IDF 方法没有考虑特征类间和类内分布,提出了一种将词频和综合评估函数值相结合的权重计算方法;最后设计了一种基于贝叶斯原理的快速分类器。实验证明该分类器简单有效。

关键词:中文文本分类;特征选择;特征权重;分类算法
DOI:10.3778/j.issn.1002-8331.2009.22.018 **文章编号**:1002-8331(2009)22-0053-03 **文献标识码**:A **中图分类号**:TP18

1 引言

文本分类是指在给定分类体系下,根据文本的内容将其分到相应的预定义的类别的过程。文本分类最初是应信息检索系统的要求出现的,随着信息网络的普及,海量的电子化文本信息的出现迫切要求由机器来自动地进行分类,这样可节约大量的人力和财力,避免人工分类带来的周期长、费用高和效率低等诸多缺陷。

文本分类过程的一般步骤:(1)预处理:将文本信息表示成计算机可以处理的结构化信息;(2)特征选择:运用特征选择算法在特征集中选择最能体现类别信息的特征,从而得出最佳的特征子集;(3)分类器训练及分类运算。其中最关键的两个步骤为"特征选择"和"分类器训练及分类运算"。

2 相关工作

(1)特征选择:近年来,在中文文本分类中经常采用的特征抽取方法包括最简单的停用词移除、互信息 MI、信息增益 IG 和 CHI 统计等。特征抽取方法的选取主要依据 Y.Yang[1]的实验结果。由于中文文本与英文文本分类相比具有相当大的差别,体现在原始特征空间的维数更大,文章表示更加稀疏,词性变化更加灵活等多个方面。因此,在英文文本分类中表现良好的特征抽取方法未必适合中文文本分类。设计特征选择的关键在于特征选择时的倾向:高频词或低频词。在以往的研究者中,都倾向于高频词,但实际上有很多低频词同样对文本分类非常关键。

(2)权重计算:有布尔权重、频度权重、TF-IDF 等,其中效果最好应用最广泛的就是 TF-IDF 权重计算公式,如式(1):

$$w_{ji}=\frac{f_{ji}\times\log(N/n_i+0.1)}{\sqrt{\sum_{i=1}^{m}[f_{ji}\times\log(N/n_i+0.1)]^2}} \quad (1)$$

其中 $w_{j,i}$ 为特征词 t_i 在文档 X_j 中的权重,f_{ji} 为特征词 t_i 在文档 X_j 中的词频,N 为文档总数,n_i 为包括词 t_i 的文档数目。

TF-IDF 是将文档集作为整体来考虑的,特别是其中 IDF 的计算,并没有考虑特征项在类间和类内的分布情况。

国内的研究还具有以下特点:一是数据集普遍偏小,一般为几千篇文本,小的只有几百篇。这使得研究结果的可信度和可比性不高;二是主要将国际上的成果应用到中文中,对中文语言的个性涉及不多;三是偏重于网页分类;四是没有可免费使用的大规模数据集。

针对以上的不足,设计了一种新的快速中文文本分类器,下面对具体的设计进行详细介绍。

作者简介:陈艳秋(1979-),女,助教,主要研究方向:人工智能、计算机网络;熊耀华(1974-),男,高级工程师,主要研究方向:软件工程、人工智能等。
收稿日期:2008-06-18 **修回日期**:2008-09-18

3 系统模型

3.1 特征选择

3.1.1 强信息特征标准

强信息特征是指那些对分类起很大作用的特征,它们应该同时具有强的分类能力和强的描述能力。强信息特征是特征选择的重点。

强信息特征包括那些只出现在个别类别文本中的中、低频特征和不均匀分布在多个类别文本中的中、高频特征,一般受以下三方面影响:

(1)频度:是最常用的特征选择测试指标,该方法认为在某一类文本中出现次数越多的特征项越能代表这类文本,因此选择在同一类文本中出现频度最高的若干特征项作为该类文本的类别特征;

(2)集中度:一个有标引价值的特征项,应该集中出现在某一类文本中,而不是均匀地分布在各类文本中;

(3)分散度:在某类文本中均匀出现的特征项对该类文本应具有较高的标引价值,若只集中出现在该类的个别文本中,而在该类的其他文本中很少出现,则该词的标引价值相对就要小多了。

显然对于某一个特征项,其频度越高、集中度越强、分散度越大,则对文本分类越有用,即分辨度越强。特征选择就是要从庞大的特征集合中找出这些强信息特征。

3.1.2 基于词频、互信息、类别信息的综合特征选择算法

(1)频度:当统计一个类别的词频时,词频最高的前几个词,基本上都是分类能力最强的词,例如统计复旦语料库中的环境类时,发现词频排在前 10 位的词如表 1。

表 1 环境类中词频前 10 的特征词

排序	特征词	词频	排序	特征词	词频
1	环境	780	6	垃圾	177
2	污染	241	7	生态	173
3	保护	210	8	城市	159
4	光华	207	9	生产	124
5	日月	207	10	环保	117

从表 1 中可以看出在这词频最高的 10 个词中,有 7 个强信息特征词,像“环境”、“污染”、“保护”、“垃圾”、“生态”、“城市”和“环保”。这说明在中文文本自动分类中,高词频对分类贡献很大。

(2)集中度:假设共有 N 个类,某特征项 t_k,t_k 与类之间的关系可能会有以下几种情况:

①特征项 t_k 只出现在一个类中,从直观上看,这个特征项非常有价值,因为可以从统计规律来确定,只要某文档中出现此特征项,就可以确认此文档的类别。

②如果 t_k 出现在两个或多个类当中,但在有些类中没有出现,那么此特征项也是有价值的。它说明了出现此特征项的文档可能会属于某些类,并且不应该属于另一些类。

③如果 t_k 在所有类中都出现了,并且出现的频率比较均匀,那么这样的特征项对分类就几乎没有价值,应当过滤掉。

也就是说,特征项出现的类别数越少,权重应该越大,可以定性地给出如图 1 所示的权重函数曲线。

根据以上权重和类别的关系,给出公式(2)来定量地表达:

$$Q_{tk}=\frac{N}{C_{tk}} \tag{2}$$

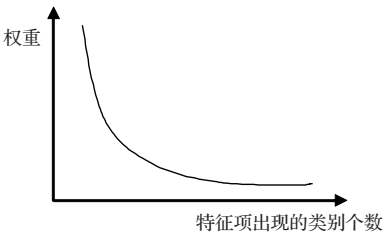

图 1 特征项出现类别个数与权重关系

其中 Q_{tk} 为特征项 t_k 的类相关系数,N 为训练文档集中包含的类别数,C_{tk} 为特征项 t_k 出现的类别个数。

(3)分散度:互信息表示特征与类别之间的相关程度。当特征的出现只依赖于某一个类别时,特征与该类型的互信息很大;当特征与类型相互独立时,互信息为 0;当特征很少在该类型文本中出现时,它们之间的互信息为负数,即负相关。频度小的特征对互信息的影响大,使得低频特征具有较大的互信息。

基于以上三个方面,提出了一种新的特征提取算法,如公式(3)所示:

$$weight(t_k,c_i)=\frac{1}{100}(tf_k\times MI(t_k,c_i)\times\frac{N}{C_{tk}}) \tag{3}$$

其中:tf_k 为特征词 t_k 在 c_i 类中出现的词频,$MI(t_k,c_i)$ 为 t_k 与 c_i 类的互信息,N 为训练文档集中包含的类别数,C_{tk} 为特征项 t_k 出现的类别个数。

在特征选择的时候,选择 $weight$ 值大的特征词。

3.2 基于词频和综合特征选择算法的权重计算

对于特征的权重,是从测试文档的角度考虑的。当一个待测文档与一个类别进行比对时,往往需要找出待测文档中出现的最能说明它属于该类的词语。此时需要考虑 2 个因素,一方面是词在 c 类中的代表性,另一方面是词在待测文章中的代表性。考虑到以上两个因素,选择词的 $weight(t_k)$作为衡量词在 c 类中代表性的标准,选择词在待测文章中出现的次数 f_k 作为词在待测文章中出现次数的衡量标准。所以采用公式(4)作为衡量权重标准。

$$value=weight(t_k)\times f_k \tag{4}$$

其中 $weight(t_k)$为特征选择时利用评估函数得到的特征评估值,代表特征与类别的相关程度,f_k 为特征词 t_k 在测试文档中出现的词频。

为了反映特征词 t_k 在测试文档中的重要程度,将测试文档划分为几个区域,如标题区域、摘要区域、正文区域等,出现在标题区域的特征项要比出现在摘要区域的特征项更能确切代表文本的内容,出现在摘要区域的特征项要比出现在正文区域中的特征项更能代表文本的内容。这样在统计出每个区域的特征项词频后得到 f_k,具体计算方法如公式(5)所示:

$$f_k=\omega_1\times f_{k1}+\omega_2\times f_{k2}+\omega_3\times f_{k3} \tag{5}$$

其中 f_{km} 为第 m 个区域的频率(m 为 1,2,3 时分别对应标题区域,摘要区域,中文区域),比例系数 $\omega_1>\omega_2>\omega_3\geqslant 1$。这样处理的目的是为了提高那些表达文本内容能力强的特征项的权重值。需要注意在训练阶段,只是简单的统计词频,并不考虑出现的位置。

3.3 改进贝叶斯的快速文本分类算法

在训练阶段,对于每一个类,分别统计在该类中出现过的词的词频、互信息值和类别信息,并根据这三个信息算出特征与类别的相关程度,然后按照 $weight$ 的大小排序,选择 $weight$

表 3　分类测试结果

类别	环境	计算机	交通	教育	经济	军事	体育	医药	艺术	政治	查全率/(%)
环境 137	123	0	2	6	2	1	0	1	0	2	89.78
计算机 193	3	172	0	10	3	0	1	0	2	2	89.12
交通 116	4	0	105	0	1	3	3	0	0	0	90.51
教育 120	0	5	0	110	1	1	0	1	1	1	91.67
经济 132	1	1	1	6	116	0	0	2	2	3	87.88
军事 150	4	0	5	4	1	123	0	3	1	9	82.00
体育 130	0	1	0	4	0	3	109	4	3	6	83.85
医药 104	2	1	0	4	2	2	0	91	2	0	87.05
艺术 208	0	0	0	10	0	0	5	0	188	5	90.38
政治 209	0	0	0	0	0	12	0	0	3	194	92.82
查准率/(%)	89.78	95.56	92.92	71.43	92.06	84.83	93.16	90.10	93.07	87.39	
*F*1/(%)	89.78	92.23	91.70	80.29	89.92	83.39	88.26	88.55	91.70	90.03	

大的前 k 个特征作为该类别的代表。注意在特征提取的时候，只是简单的统计在某类文档中出现的频率，并不考虑出现的位置。

在测试阶段，将一篇待分类文档分词后，按照词出现的位置(标题区域、摘要区域、正文区域)的不同，根据公式(5)计算词频；然后与第一个类别提取出的 k 个词进行比对，若出现则按照公式(4)计算权值，最后将得到的权值相加，作为测试文档与该类比较的最终结果。待测试文档与所有类别比较完毕后，对最终结果由大到小排序，选出结果最大的作为最终的分类结果。

4　实验结果

4.1　实验数据

从高校 BBS 的 10 个版面收集了 10 类共 2 815 篇中文文档作为实验数据，各文档分布如表 2。

表 2　实验文档

类别	文档数	所占比例	类别	文档数	所占比例
计算机	200	0.071 0	经济	325	0.115 5
艺术	248	0.088 1	医药	204	0.072 5
教育	220	0.078 2	军事	249	0.088 5
交通	214	0.076 0	政治	505	0.179 4
环境	200	0.071 0	体育	450	0.159 9

4.2　评价指标

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：

$$查准率(precision)=\frac{分类的正确文本数}{实际分类的文本数}$$

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：

$$查全率(recall)=\frac{分类的正确文本数}{应有的文本数}$$

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测试值，其数学公式如下：

$$F1\ 测试值=\frac{准确率\times查全率\times2}{准确率+查全率}$$

4.3　开放测试结果

收集到复旦的语料库作为测试样本，复旦预料库将文本分成 20 大类，只挑选与本设计的分类器相关的 10 个类别，具体为：环境、计算机、交通、教育、经济、军事、体育、医药、艺术和政治。由于复旦预料库中的类文本的数量较大，从中随机挑选了部分文本，在关键字个数为 20 的情况下进行测试，结果如表 3。

从表中可以看出，测试结果的准确程度非常高，这说明所设计的分类器是非常成功的。

为了考察关键字数对分类性能的影响，对分类器在不同关键字数下的性能进行了测试，下面列出环境类的分类效果，如表 4 所示。

表 4　环境类在不同关键字数下测试结果

关键字数	查全率 *Re*	查准率 *Pr*	关键字数	查全率 *Re*	查准率 *Pr*
10	0.876	0.853	200	0.925	0.923
20	0.898	0.879	250	0.919	0.909
50	0.909	0.881	300	0.892	0.898
100	0.911	0.914	500	0.865	0.853
150	0.923	0.920	800	0.864	0.874

从中可以看到，准确度(包括查全率和查准率)在关键字数为 100 以前急剧上升，然后缓慢增加，在选取的关键字数为 200 左右时分类效果达到最好，最后随着关键字的增加准确率缓慢下降。

5　结论

为了提高中文文本分类的效率与精度，从特征提取，权重计算，分类算法入手，设计了一种新型的分类器。该分类器能够选择出大部分对分类最有效的特征词，利用这些强信息特征词，配合分类算法取得了比较满意的分类结果。

参考文献：

[1] Yang Y，Pedersen J P.A comparative study on feature selection in text categorization[C]//Proceedings of the 14th International Conference on Machine Learning，1997.

[2] 李荣陆.文本分类及其相关技术研究[D].上海：复旦大学，2005.

[3] 薛得军.中文文本自动分类中的关键问题研究[D].北京：清华大学，2004.

[4] 陈治纲.基于向量空间模型的文本分类系统研究与实现[D].天津：天津大学，2005.

[5] 刘东绪.在自然汉语中进行分词和词性标注[D].成都：电子科技大学，2003.