一种基于 VSM 的中文网页分类方法*

孔令成,郑 诚,吴永俊
(安徽大学 计算智能与信号处理重点实验室,安徽 合肥 230039)

摘 要:本文应用有指导机器学习方法实现了一个分类器。运用改进型的 MI 进行特征提取,并对传统的 TFIDF 加权公式进行了改进。实验结果表明该分类器有较高的分类质量,满足了中文网页自动分类的要求。

关键词:网页分类;文本;算法;特征
中图分类号:TP391.1 **文献标识码**:A

Chinese Web-page classification algorithm based on VSM

KONG Ling Cheng, ZHENG Cheng, WU Yong Jun
(Key Lab. of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: Web-page classification plays an important role in data mining, and it is one of the key topics in information retrieval. This paper makes use of supervised machine learning theory to implement a Web-page classifier. The MI is used improved for feature extraction and improve traditional TFIDF formula. The experiment results show that the classifier is feasible and effective.

Key words: Web-page classification; text; algorithm; feature

网络的迅速发展,使人们不仅面临信息爆炸,同时也面临着如何从浩如烟海的信息中获取自己所需信息的难题。如何有效地组织和处理海量的信息,并过滤和管理网络资源,已成为必须面对的问题。

为了网页信息的有效组织和检索,人们开发了各种网络信息搜索器(比如搜索引擎),在一定程度上确实提高了网络信息的利用率。与文本分类技术相比较,网页分类更加复杂,这是由网页的结构特征决定的,但是网页的信息主要是通过文本的方式向人们传递的,所以在对网页分类之前,首先要对其中的文本进行提取,对所提取的文本分类,最终使网页分类问题转化为文本分类问题。

目前,文本分类技术的研究比较活跃,已经出现了多种文本分类算法,并且被广泛应用于多个领域:信息检索、搜索引擎、文本数据库等。文本分类算法[1-3]基本是基于概率统计模型,例如贝叶斯分类算法(Naive Bayes),支持向量机(SVM)、K 近邻算法(KNN)等等。本文就是基于互信息(MI)提出一种改进的特征提取方法,并根据 TFIDF 提出一种新的特征权值计算方法构建中文网页分类器。实验表明,改进后的特征提取和特征权值计算算法在中文网页分类过程中取得了良好的效果。

1 网页预处理

网页分类之前首先要进行预处理,实际上就是 HTML 解析,把解析出来的内容用于文本分类,选取网页中的下面这些文本用于分类:

(1)锚文本。锚文本是网页中用于指示所连接网页内容的提示,由于后面要对提取的文本进行分类,所以只提取文字形式的锚文本。

(2)title 文本。这样的文本可能是网页中最重要的标签,必须取得。

(3)meta 标签。其重要的功能就是设置关键字,网页的制作者往往都设置了关键字,来提高网页的搜索点击率。可以利用 meta 标签中的有关文本内容进行网页分类。

(4)主文本。上面这些信息获取之后,网页中剩余的文本信息还在各种 HTML 标签中,在 HTML 源文件中,主文本有可能不是连续出现的。主文本一般是网页中文字最集中的较长的字符串,查看源文件,那些比较长的

* 基金项目:安徽省高等学校省级自然科学研究重点项目(KJ2009A57)

字符串是整个出现在1个标签中的,因此提取出标签中的文本,并比较长度,选择较长的某几个作为主文本,利用它们进行分类。

网页中像jave script和sytle这样的信息,如果把这些信息带到后面的信息提取中,会使所获取的文本准确度大大地降低,所以必须在网页中获取相关文本前就除掉。

文本首先要确定的问题就是表示文本的基本单位,用于表示文本的基本单位通常称为文本的特征或特征项。中文文本不同于英文文本,英文文本以空格为分隔符,非常明确。而中文文本需要对其进行分词处理才能得出每个特征。本文采用中科院计算技术研究所汉语词法分析系统ICTCLAS3.0[4]进行分词。对于文本中的特征项,能标识文本特性的往往是文本中的实词,如名词、动词等。而文本中的一些虚词(如感叹词、介词等),对于标识文本的类别特性并没有多少贡献。如果把这些对文本分类没有意义的虚词作为特征,将会带来很大噪音,降低文本分类的效率和准确率。因此,在提取文本特征时,应首先考虑剔除这些对文本分类没有用处的虚词,而在实词中,又以名词和动词对于文本的类别特性的表现力最强,所以只保留那些对于文本分类有用的实词即:名词、动词。即便剔除了文本中的虚词,要对文本分类的数据量仍然会很大,为了进一步减少影响文本分类的噪音,则需要提取出对文本分类贡献大的特征项。

2 特征提取

特征提取就是提取出最能代表某篇文章或某类的特征项,以达到降维的效果从而减少文本分类的计算量。典型特征提取方法:信息增益(Information Gain),互信息(MI)、文档频度(DF)。传统的MI特征提取方法:

$$MI(t,c)=P(t,c)\left|\log P(t|c)-\log P(t)\right| \tag{1}$$

其中:$P(t,c)$表示t和c的同现概率,$P(t|c)$表示t在类别为c的文本中出现的概率,$P(t)$表示t在所有文本中的出现概率。本文使用改进型的MI方法:

$$MI(t,c)=P(t,c)\left|\log P(t|c)-\log P(t)\right|\log(n/nf) \tag{2}$$

其中:$P(t|c)$、$P(t)$、$P(t,c)$的含义同上,nf为出现t的文档数,n为训练集中的所有文档数。它基于如下的假设:如果词条出现的文档数接近训练集中所有的文档数时,即n趋向于nf时,$\log(n/nf)$趋向于0,此类词条应该过滤掉,并且适当地提高低频词的权重。这样计算某个特征词可能会出现在几个类中,为使其应用于多类中,可取其均值,即:

$$MI(t)=\frac{\sum_{i=1}^{m}MI(t,c)}{m} \tag{3}$$

计算出所有特征词的统计值后,从大到小进行排序,然后根据需要从上到下选取一定数量的特征词构建文本分类的特征词库。

3 特征加权及向量化

TFIDF算法及其改进型[5]有多种公式,本文使用一种新的改进的TF-IDF公式来计算特征词的权重。TF-IDF公式有很多变种,比较常见的TF-IDF公式:

$$w(t_i,\bar{c}_j)=\frac{tf(t_i,\bar{c}_j)\times\log(N/n_i+0.1)}{\sqrt{\sum_{t=\bar{c}j}\left[tf(t_i,\bar{c}_j)\times\log(N/n_i+0.1)\right]^2}} \tag{4}$$

首先把NC定义为类别个数,上式中j的取值范围是$(1,2,\cdots,NC)$,N为所有文档数目,n_i为含有词条t_i的文档数目。$tf(t_i,\bar{c}_j)$表示为第i个特征项t_i在第j类$\bar{c}_j$上的平均词频。

$$tf(t_i,\bar{c}_j)=\frac{\sum_k \omega_{ijk}}{\left|\bar{c}_j\right|} \tag{5}$$

其中,ω_{ijk}是特征项t_i在$\bar{c}_j$类中第k篇文档中的词频,k的取值范围是:$(1,2,\cdots,\left|\bar{c}_j\right|)$。根据TF-IDF公式,文档集中包含某一词条的文档越多,说明它区分文档类别属性的能力越低,其权值越小。另一方面,某一文档中某一词条出现的频率越高,说明它区分文档内容属性的能力越强,其权值越大。

网页不同于一般的文本,页面中包含了诸如<head>,<title>、<body>、<meta>等标记用于描述页面的标题,主体,关键词等tag信息,根据它们所包含的分类信息定义不同的权值,具体如表1所示。

表1 网页标签及权值

标签	权重	标签	权重
Title	5	Head	4
H1	4	Font size>=7	4
H2	3	Font size=6	3
H3	2	Font size=5	2
H4	1	Font size=4	2
H5	1	Font size<=3	1
Strong	3	Font size=1	2

针对网页的特征,对于ω_{ijk}的计算,可以表示成标题权重与出现次数乘积的代数和,如果特征t_i被多个标记修饰(假设有k个),设k个的权重分别是$w1,w2,\cdots,wk$,对应的ti出现的次数分别是$c1,c2,\cdots,ck$,那么:

$$\omega_{ijk}=\sum_{i=1}^{k}w_i\,c_i \tag{6}$$

特征项的类间分布信息用式(7)来表示:

$$E(t_i)=\frac{\left[tf(t_i,\bar{c}_j)-\bar{X}\right]^2}{\bar{X}} \tag{7}$$

其中,$\bar{X}=\frac{1}{NC}\sum_{i=1}^{NC}tf(t_i,\bar{c}_j)$,$j=1,2,\cdots,NC$。

最终改进后的特征权值计算公式为:

$$w(t_i,\bar{c}_j)=E(t_i)\times\frac{tf(t_i,\bar{c}_j)\times\log(N/n_i+0.1)}{\sqrt{\sum_{t\in\bar{c}}\left[tf(t_i,\bar{c}_j)\times\log(N/n_i+0.1)\right]^2}} \quad (8)$$

这样的改进考虑到特征在类间的分布信息。

向量空间模型(VSM)[6]是信息检索领域应用广泛且效果较好的模型,在该模型中,文档被看成一系列无序词条即特征项的集合,对每个特征项加上1个对应的权值,把文档映射成1个向量。计算机不能直接处理文本文件,基于向量空间模型的思想将中文文本信息以被计算机容易处理的向量的形式表示为:$D_i=(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$,其中$D_i$为某一文本,$t_i$为有意义的特征词,$w_i$为特征词对应的权值,$n$表示特征向量空间的维数。

4 实验结果与分析

实验数据来源于新浪,搜狐等各大网站。共下载了7 000个网页,分为军事、经济、体育、教育、计算机、环境、艺术7类,每个类包含1 000个网页,其中800个网页当做训练集,另外200个网页当做测试集。

实验使用KNN作为分类器,主要是因为KNN作为一种传统的模式识别方法用于分类时,在查准率和查全率上的表现令人满意。KNN在已知类别样本中寻找与待分类样本最相近的K个样本,样本之间的相似度可以通过向量之间的余弦来度量,如:

$$sim(D_i,C_j)=\frac{\sum_{k=1}^{m}W_{ik}*W_{jk}}{\sqrt{\sum_{k=1}^{m}W_{ik}^2}\sqrt{\sum_{k=1}^{m}W_{jk}^2}} \quad (9)$$

其中,D_i为待分类文本的特征向量,C_j为第j类的类别特征向量,m为特征向量的维数,W_k为向量的第k维。

KNN分类基于,未知类别的样本预测为在K个最近邻样本中含有最多实例的类别。

在利用改进后的算法进行试验的同时,比较算法改进前后的效果,从原始的MI和TFIDF公式分别进行特征提取和特征权值计算,并在相同的数据训练集和数据测试集上做试验。

为了评价实验的结果,采用较为通用的性能评价方法,即查准率和查全率。

查准率是分类器在某类别中做出的正确分类个数与分类器在该类别上做出的所有分类个数的百分比。

查全率是分类器在某类别中做出的正确分类个数与该类实际应有文本个数的百分比。

表2是算法改进前后的试验数据比较:

表2 算法改进前后的试验数据比较

测试数据集	改进前		改进后	
	查准率/%	查全率/%	查准率/%	查全率/%
计算机	81.5	85.5	84.2	87.0
军事	79.3	84.0	82.6	86.0
环境	76.3	81.5	80.6	85.0
经济	77.7	80.5	81.2	84.0
教育	75.6	81.0	81.8	83.5
艺术	72.8	77.0	78.7	83.0
体育	79.9	83.5	82.9	86.0
平均	77.6	81.9	82.7	84.9

由于下载的网页是从网站中寻找类别信息,而且网页内容主要以中文文本为主,所以导致查全率和查准率普遍不高,假如实验用的语料库是标准的语料库,查全率和查准率应该会更高。

对于不同的类别,查全率和查准率也会不同,计算机、军事和体育由于类别信息比较明显,所以查全率和查准率相比其他类都比较高,而艺术类由于包含的信息广泛,且模糊,所以查全率和查准率相对较低。

新的算法主要是改进了原始算法对网页的类别贡献比较大的低频词却容易被删除的缺点,比如1个网页中出现了某个体育明星的名字,属于体育类的可能性就比较大,但是这个体育明星的名字在一般体育类网页中的出现几率还是比较小的,那么体育明星的名字就是对体育类网页的类别贡献比较大的低频词,用原始算法就难以分出类别,改进后的算法相对容易分出来。

从表2可以看出来,利用改进后的MI和TFIDF进行特征提取和特征权值计算,查全率和查准率相对于算法改进前都有比较大的提高。

本文使用KNN对于改进后的MI和TFIDF进行了测试,结果表明查全率和查准率相对于原始算法都有较大提高,从而证明了在网页分类上的可行性。同时特征向量维数,训练语料库的选取和大小,以及分类器的选择也会对分类结果产生影响。

参考文献

[1] 原福永,于歌,崔春华.基于特征选择的网页分类方法研究[J].计算机工程与设计,2007,28(17).

[2] 柴玉梅,王宇.基于TFIDF的文本特征选择方法[J].微计算机信息,2006,22(8-3).

[3] YANG Yi Ming, LIU Xin. A re-examination of text categorization methods [C]. Proceedings of SIGIR-99,22nd ACM International Conference on Research and Development in Information Retrieval.1999:42-49.

[4] 中科院汉语词法分析系统:http://www.i3s.ac.cn,2008.3.18.

[5] 张玉芳,彭时名,吕佳.基于文本分类TFIDF方法的改进与应用[J].计算机工程,2006,32(19).

[6] 郭庆琳,李艳梅,唐琦.基于VSM的文本相似度计算的研究[J].计算机应用研究,2008,25(11).

(收稿日期:2009-05-12)