一种基于反向文本频率互信息的文本挖掘算法研究

周 戈
(重庆青年职业技术学院, 重庆 400712)

摘 要: 针对传统的文本分类算法存在着各特征词对分类结果的影响相同,分类准确率较低,同时造成了算法时间复杂度的增加,在分析了文本分类系统的一般模型,以及在应用了互信息量的特征提取方法提取特征项的基础上,提出一种基于反向文本频率互信息熵文本分类算法。该算法首先采用基于向量空间模型(vector space model, VSM)对文本样本向量进行特征提取;然后对文本信息提取关键词集,筛选文本中的关键词,采用互信息来表示并计算词汇与文档分类相关度;最后计算关键词在文档中的权重。实验结果表明了提出的改进算法与传统的分类算法相比,具有较高的运算速度和较强的非线性映射能力,在收敛速度和准确程度上也有更好的分类效果。

关键词: 文本挖掘; 互信息; 向量空间模型; 权重
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0487-03
doi:10.3969/j.issn.1001-3695.2012.02.023

Reverse text frequency based on mutual information on text categorization

ZHOU Ge
(*Chongqing Youth Vocational & Technical College*, *Chongqing* 400712, *China*)

Abstract: In view of the traditional text classification algorithm has the characteristics of classification results on the influence of the same, the classification accuracy rate is low, caused at the same time algorithm time complexity increases, based on the analysis of the text classification system of the general model, as well as in the application of mutual information feature extraction method based on feature, this paper put forward a method based on reverse text frequency mutual information entropy text classification algorithm. The algorithm first used based on the VSM on the text sample vector feature extraction, then the text imaged to extract keywords set, selection of key words in the text, using mutual information to represent and computational lexicon and document classification correlation, finally calculated key words in the document weight. The experimental results show that the proposed algorithm and the traditional classification algorithm, has high computing speed and strong nonlinear mapping ability, the speed of convergence and accuracy are better classification effect.

Key words: text categorization; mutual information; vector space model; weight

0 引言

文本分类是中文信息处理的一个重要的研究领域。其目标是在分析文本内容的基础上,给文本分配一个或多个比较合适的类别,从而提高文本检索、存储等应用的处理效率。目前国内外基于内容文本信息过滤的研究一般集中在分类的核心算法上,主要可概括为用户模板的构建及算法研究和与文本的匹配技术两个方面,这也是文本信息过滤的两大关键技术所在。

目前已有大量的统计分类技术应用于文本自动分类中,这些传统的分类方法有着不同的缺陷。传统的聚类方法具有较好的分类精度,但是其训练时间较长[1];向量距离分类法有算法简单、分类速度快等特点,但是过于依赖经过平均运算得到的中心向量,因而分类精度不高[2];传统的贝叶斯对文本分类规则的判别具有独特的优势,但文本特征维数过高导致神经网络不易收敛、学习时间太长[3]。所以,目前很难找到一个执行效率高,精确率和召回率都很理想的算法[4]。

针对传统的文本分类算法存在着各个特征词对分类的结果影响相同,分类准确率较低,同时造成了算法时间复杂度的增加,本文在分析了文本分类系统的一般模型,以及在应用了互信息量的特征提取方法提取特征项的基础上,提出了一种基于反向文本频率互信息熵文本分类算法。

1 文本特征提取分类

文本特征提取是进行文本分类训练和识别的基础。本文采用算法的基本思路是首先基于向量空间模型(VSM),即把一篇文本视为 N 为空间中的一个点。点的各维数据表示该文档的一个特征(数字化的特征)。而文档的特征一般采用关键词集,即根据一组预定义的关键词,以某种方法计算这些关键词在当前文档中的权重,然后用这些权重形成一个数字向量,这就是该文档的特征向量。本文首先需要解决的是提取文本中的关键词集,再计算该关键词在文本中的权重,然后对其进

收稿日期: 2011-07-29; **修回日期**: 2011-09-05
作者简介: 周戈(1978-),男,重庆人,讲师,硕士,主要研究方向为数据仓库与数据挖掘(4096185@qq.com).

行分类。

1.1　文本关键词集提取

提取关键词的最终目的是为了对文本进行有效的分类。对于一些词,例如“的”,对应文本分类不可能有任何帮助,或者“计算机”一词对进行“科普类”和“小说类”文章的分类也没有任何帮助。所以说关键词集是与分类目标相关的。本文在提取关键词集中有两个步骤:

首先根据词汇与预定义分类文本的相关程度来筛选关键词。使用一个训练文档集(其中各文档的分类已经由人工指定),通过计算其中词汇与文档分类的相关程度,选择相关程度高的词汇作为表达文档特征的关键词。如果词汇 w 在 C_i 类文本中出现的频率很高,就用它作为一个关键词:

$$tf(w,C_i)=\frac{\text{count}(w|C_i)}{\text{count}(w'|C_i)} \tag{1}$$

其中:$\text{count}(w|C_i)$ 表示在 C_i 类文档中 w 出现的总次数;$\text{count}(w'|C_i)$ 表示 C_i 类文档中的总词汇数。计算 C_i 类文档中各词汇的词汇频率后,设定一个阈值,选择大于该阈值的词汇作为 C_i 类的关键词。将各类关键词集合并后,形成整个系统的关键词集。但是该方法存在一定的局限性,即对规模较大的文本来说,难以形成整个系统的关键词集,本文采用的是基于反向文档频率互信息熵来对文档的关键词集进行提取。

文档频率是指词汇 w 在整个文档集中的文档频率,而式(2)是指在类 C_i 子集中的文档频率。因而这里的文档频率的计算为

$$DF(w,C_i)=\frac{n_w}{N} \tag{2}$$

其中:n_w 是包含 w 的文档总数,N 是总文档数。

词汇 w 的反向文档频率计算方法为

$$\text{TFIDF}(w,C_i)=tf(w,C_i)\times\log(1/DF(w,C_i))=tf(w,C_i)\times\log(N/n_w) \tag{3}$$

计算 C_i 类文档中各词汇的 TFIDF 后,设定一个阈值,选择大于该阈值的词汇作为 C_i 类的关键词。

互信息指标是用于表示两个特征共同出现的程度。在这里,如果词汇 W 和类 C 总是共同出现,那么它们的互信息度高,W 就是 C 类文档的一个特征词。

$$MI(w,C_i)=\log\left(\frac{P(w,C_i)}{P(w)P(C_i)}\right)=\log\left(\frac{P(C_i|w)}{P(C_i)}\right) \tag{4}$$

其中:$P(w)$ 是在整个训练集中出现词汇 w 的文档概率(用频率代替);$P(C_i)$ 是在训练集中属于类 C_i 的文档概率;$P(w,C_i)$ 表示在训练集中既出现 w 又属于类 C_i 的文档概率。

此外,w 与 C_i 的互信息度高,并不说明 w 与另一个类 C_j 的互信息度就一定低。为了更好地区分两个类,本文选择仅与一个类的互信息度高的词汇。但这种表达是很理想化的。实际上可以选择那些与不同类的互信息度差距较大的词汇作为关键词。表示这一特征的方法是求词汇 w 的互信息度的均方差:

$$\sigma(w)=\sqrt{\sum_{i=1}^{m}(MI(w,C_i)-MI_{\text{avg}}(w))^2} \tag{5}$$

其中:$MI_{\text{avg}}(w)$ 为 w 的平均互信息度,其公式表示为

$$MI_{\text{avg}}(w)=\sum_{i=1}^{m}P(C_i)\times MI(w,C_i)$$

互信息的一个缺点是没有考虑 w 在某类文档中的词汇频率,因而稀有词汇常常可以有很大的权重。本文提出的算法是:

$$MI(w,C_i)=\log\left(\frac{P(w,C_i)}{P(w)P(C_i)}\times TF(w,C_i)\right) \tag{6}$$

其中:$TF(w,C_i)$ 是词汇 w 的词频在 C_i 类文章中的词汇频率:

$$TF(w,C_i)=\frac{\text{count}(w|C_i)}{\text{count}(w)} \tag{7}$$

其中:$\text{count}(w)$ 是 w 在所有文章中出现的词汇数,$\text{count}(w|C_i)$ 是 w 在 C_i 类文章中出现的词汇数。计算各词汇与 C_i 类的互信息度后,设定一个阈值,选择大于该阈值的词汇作为 C_i 类的关键词。将各类的关键词集合并后,形成整个系统的关键词集。

1.2　关键词权重计算

在提取特征词中,笔者希望取 w 为特征词,并根据是否包含 w 将整个文本集分为两个子集后,各类文本在两个子集内部分布得非常不均匀。理想的情况是,正好一个子集包含一个类。这一两个子集内部的熵就非常小,而整个系统的熵是两个子集熵的和,因而也会变小。这样,根据 w 划分子集后,系统就产生了一个熵增益(实际上是熵减)。通过比较不同词汇对系统产生的熵增,选择那些熵增很大的词汇作为关键词。

使用 w 划分子集前,整个系统的熵(entropy)为

$$E=\sum_{i=1}^{m}P(C_i)\log(1/P(C_i)) \tag{8}$$

其中:$P(C_i)$ 为文本集中 C_i 类文本出现的概率(频率)。划分后,系统的熵为

$$E^w=\sum_{i=1}^{m}P(C_i|w)\log(1/P(C_i|w))+\sum_{i=1}^{m}P(C_i|\bar{w})\log(1/P(C_i|\bar{w})) \tag{9}$$

其中:$P(C_i|w)$ 是在包含词汇 w 的文本子集中 C_i 类文本出现的概率;$P(C_i|\bar{w})$ 则是在不包含词汇 w 的文本子集中 C_i 类文本出现的概率。根据以上两个公式,使用 w 作为关键词的熵增为

$$G^w=E-E^w \tag{10}$$

根据前面提取的一组关键词,表示为$\langle K_1, K_2, \cdots, K_n\rangle$,需要将任意一篇文档转换为数字向量,如$\langle q_1, q_2, \cdots, q_n\rangle$,其中,$q_i$ 是关键词 K_i 对于当前文档的权重,即重要性。计算某个关键词对一篇文本的权重,本文采用的是以关键词的反向文档频率作为其权重。

TF-IDF 判断关键词对于文档的重要性时,不仅考虑一个关键词在文档中出现的频率(即上述的词频),而且考虑该关键词在所有文档中出现的频率(即文档频率)。如果一个关键词在很多文档中都出现,那么它对于当前文档的重要性就比较低。关键词 t_i 对于文档 d 的 TF-IDF 权重计算的方法是:

$$\text{TFIDF}(t_i,d)=\frac{tf(t_i,d)\times\log(N/n_{t_i}+0.01)}{\sqrt{\sum_j[tf(t_j,d)\times\log(N/n_{t_j}+0.01)]^2}} \tag{11}$$

其中:N 表示文本总数;n_t 表示出现关键词 t 的文本数;N/n_t 称为 t 的反向文档频率。式(11)的分子中综合了 t 的词频和反向文档频率两个因素,因而能够更好地反映 t 与文档 d 的关系;式(11)的分母中[ ]内部分形式上与分子相同,但其中的 t_j 是指各个关键词,表示求所有关键词的平方和,其目的是归一化关键词的权重。

在实现中,本文使用训练文档集作为计算 TFIDF 的基础:N 表示训练文档集中的文本总数,n_t 表示关键词 t 在训练文本集中出现的文本数。当对新的文本进行编码时,对任意一个关键

词 t,只需要统计它在该文本中的词频(tf),结合在训练集中已统计的 N 和 n_t,就可以计算出 t 的 TFIDF 权重。最后本文还考虑了文档中一些停用词,即各种文档中都经常出现的、不能反映文档内容特征的常用词,如助词、语气词等(已有停用词表)。本文采用的是反向文档频率必须要考虑停用词,因为排除可以提高筛选关键词的效率。在筛选关键词前,首先排除停用词。

2　实验结果与分析

2.1　数据来源

本文主要做了两组实验:

第一组数据来自于历史方面的文章,主要是通过 spider 从互联网上搜集而来的,其中包含了 502 篇文章。首先经过本文提出的特征提取算法处理去掉停用词,之后保留了 3 021 个特征关键词,通过人工分类,将所有的文档分为六大类,如表 1 所示。

表 1　历史文档分类

	古代史	近代史	时事	外国古史	外国近史	外国时事
篇数	135	100	78	59	80	50

第二组数据是互联网上的文学类文章,同样是通过互联网上搜集得到的,同样去掉停用词后一共保留了 29 122 个文本关键词,这里还要去掉所有的文章中出现的频率过小或者过大的词。一般是小于 0.05% 或者大于 95% 的词要过滤掉,最后剩下关键词有 8 329 个,同样经过人工分工可以分为以下不同的六类,如表 2 所示。

表 2　文学文档分类

	散文	诗歌	小说	古典文学	外国文学	戏剧
篇数	135	100	78	59	80	50

文学类数据中各个类别的区分不像其他类别比较明显,因为文学涵盖的范围非常大,本文给出的是为了实验而选择的一部分。

本文采用了传统的查全率和查准率来衡量提出算法的有效性能和实际应用价值。查全率主要是衡量在特定的文档类别中正确分类的文档数目,以及在所有的文档库中属于该类别的文档数目的比率。查准率主要是某个文本类别中所有的正确分类的文档数与这个类别的文档总数之间(自动分类)的比率,主要用来分析分类算法的查准率。而对于一个特定的分类算法来说,查全率和查准率都很难做到两全其美:如果查全率越高,那么查准率就会受到一定的影响;如果查准率越高,那么查全率就会降低。两个指标定义如下:

查准率 = (检索出的相关信息量/检索出的信息总量) × 100%

查全率 = (检索出的相关信息量/系统中的相关信息总量) × 100%

本文采用平均准确度来衡量本文分类算法的标准。平均准确率主要是通过评价任意不同的两个文档之间的分类效果来得到,定义如下:

$$\text{平均准确率} = \frac{\text{查全率} + \text{查准率}}{2}$$

2.2　实验设置分析

本文所有的实验都是在 PC P4 T2310 1.86 GHz CPU,2 GB RAM,Intel 82865G 显卡的计算机上进行的,实验环境为 MATLAB 7.0。为了验证本文算法具有一定的优越性,对比了传统的文本分类算法,并在仿真平台上进行仿真对比。仿真实验结果如图 1 和 2 所示。

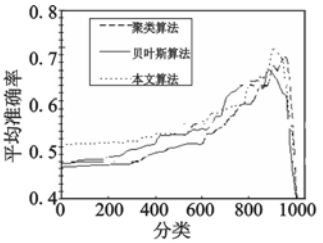

图1　历史类文章平均准确率对比

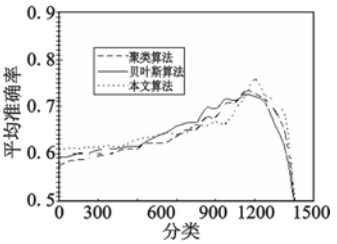

图2　文学类文本聚类准确率对比

从图 1 和 2 中可以看出,本文提出的算法能够有效地对文本进行特征提取并加以分类。通过对比传统的聚类算法分类以及贝叶斯文本算法分类可以发现,本文算法在平均准确率上要明显高于其他两种算法,算法的分类准确率获得了明显的提升。

3　结束语

文本分类是指根据文档的内容或属性,将大量的文本归到一个或多个类别的过程。它在日常生活中具有非常重要的意义。本文在分析了文本分类系统的一般模型,以及在应用了互信息量的特征提取方法提取特征项的基础上,提出了一种基于反向文本频率互信息熵文本分类算法。该算法首先采用基于向量空间模型对文本样本向量进行特征提取;然后对文本图像提取关键词集,筛选文本中的关键词,采用互信息来表示并计算词汇与文档分类的相关度;最后计算关键词在文档中的权重。实验结果表明了提出的改进算法与传统的分类算法相比,具有较高的运算速度和较强的非线性映射能力,在收敛速度和准确程度上也有更好的分类效果。

参考文献:

[1] HUANG J Z X, NG M K, RONG Hong-qiang, *et al*. Automated variable weighting in K-means type clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[2] 刘斌,黄铁军,程军,等.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.

[3] 朱明,王俊普,蔡庆生.一种最优特征集的选择算法[J].计算机研究与发展,1998(9):803-805.

[4] 胡佳妮,徐蔚然,郭军,等.中文文本分类中的特征选择算法研究[J].光通信研究,2005(3):44-46.

[5] 潘有能.一个自动分词分类系统的实现[J].情报学报,2002,21(1):38-41.

[6] 邓乃扬,田英杰.数据挖掘中的新方法:支持向量机[M].北京:科学出版社,2004.

[7] 巩知乐,张德贤,胡明明.一种改进的支持向量机的文本分类算法[J].计算机学报,2009,26(7):165-168.

[8] 解冲锋,李星.基于序列的文本自动分类算法[J].软件学报,2011,13(4):783-788.

[9] 韩家炜,孟小峰,王静,等.Web 挖掘研究[J].计算机研究与发展,2011,38(4):405-411.