第16卷 第3期
2006年3月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.3
Mar. 2006

Web文本分类技术研究及其实现

饶文碧,柯慧燕
(武汉理工大学 计算机学院,湖北 武汉 430070)

摘 要:随着Internet的飞速发展,Web文本分类研究已经得到了人们密切的关注,并取得了大量的研究成果。文中讨论了Web文本分类过程中的几个关键技术;针对传统的Web文本分类方法缺乏认知自主性和不能再学习的特点,提出了一种扩展的Web文本分类模型和算法。通过系列实验表明,该算法具有较高的分类精度和查准率。

关键词:Web文本分类;向量空间模型;特征提取;反馈判定
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1005-3751(2006)03-0116-03

Research and Implementation of Web Text Classification

RAO Wen-bi, KE Hui-yan
(School of Computer Science, Wuhan University of Technology, Wuhan 430070, China)

Abstract: With the development of Internet at full speed, the research of Web text classification has already got people's close concern. A large amount of research results have been got. This paper has discussed several key technologies in the course of Web text classification in detail at first; Then directing against the traditional classification algorithm of Web text lack of cognitive independence and studying again, it proposes an extended Web text classification model and algorithm. Through a series of experiments, can get the result that such algorithm has higher classification precision and recall.

Key words: Web text classification; vector space model; feature extraction; feedback and judge

0 引 言

随着Internet的快速发展,互联网上出现了海量的、异质的Web信息资源,其中Web文本信息占了主要地位。如何从浩瀚的Web文本信息资源中准确获取所需信息,已成为信息处理的一个关键问题。于是人们将数据挖掘技术应用到Web的知识发现中形成了现在的Web挖掘技术。其中作为Web挖掘技术的关键部分,Web文本分类技术已经得到了人们的广泛关注。

Web文本分类技术是一种典型的有教师的机器学习方法,作为Web文本挖掘的一项重要技术,它是指将Web文档集合中每个文档归入一个预先定义的类别之中。目前Web文本分类算法及应用已有大量的研究,其中涌现出的Web文本分类算法很多。关于文本分类方法主要有朴素贝叶斯(Naive Bayes)方法[1]、决策树(Decision Tree)和k最近邻参照(kNN)[2]以及基于粗糙集的分类方法[3]等等。

文中首先对文本分类过程中的关键技术:文本的表示、文本预处理(构造词典、分词)、特征提取、训练算法、分类算法进行了详细介绍,然后提出了一种扩展的基于向量空间模型的Web文本分类模型和算法。

收稿日期:2005-06-01
作者简介:饶文碧(1967—),女,湖北武汉人,教授,主要研究方向为人工智能、数据挖掘。

1 Web文本分类中的关键技术

1.1 文本的表示

计算机没有类似人类的智能,人阅读完文章后可以产生自身对文章的理解,而计算机却没有这样的能力。为了便于计算机的处理,文本必须表示为计算机可以识别的格式。

目前文本的表示模型有多种:布尔逻辑型、向量空间型(VSM)、概率型以及混合型等。文中选取了信息处理领域常用的向量空间模型。在向量空间模型中,文本泛指各种机器可读的记录,用 D(Document)表示,特征项(Term,用 t 表示)是指出现在文档 D 中且能够代表该文档内容的基本语言单位,主要是由词或者短语构成,文本可以用特征项集表示为 $D(T_1,T_2,\cdots,T_n)$,其中 T_k 是特征项,$1\leqslant k\leqslant N$。对含有 n 个特征项的文本而言,通常会给每个特征项赋予一定的权重表示其重要程度,即 $D=D(T_1,W_1;T_2,W_2;\cdots,T_n,W_n)$,其中 W_k 是 T_k 的权重,$1\leqslant k\leqslant N$。关于权重的计算,文中选取了常用的TF-IDF公式[4]:

$$W(t,d)=\frac{(1+\log_2 tf(t,d))\times\log_2(N/n_t)}{\sqrt{\sum_{t\in d}[1+\log_2 tf(t,d))\times\log_2(N/n_t)]^2}} \quad (1)$$

其中,$W(t,d)$为词 t 在文本 d 中的权重,而 $tf(t,d)$为词 t 在文本 d 中的词频,N 为训练文本的总数,n_t 为训练文本集中出现 t 的文本数,分母为归一化因子。

1.2 文本预处理(构造词典、分词)

Web文本作为一种非结构化的数据类型,其特点表现为特征空间的高维性、文本特征表示向量的稀疏性及文本主题特征表现不突出等特点。与数据库和数据仓库中的结构化数据相比,Web文本具有有限的结构或者根本就没有结构。文本信息源的这些特征使得现有的数据挖掘技术无法直接应用于其上,因此需要对Web文本进行预处理,抽取其特征并用结构化的形式保存,作为文本的中间表示形式。

在对文档进行特征提取前,需要先进行文本信息的预处理,对英文而言需进行Stemming处理,中文的情况则不同,因为中文词与词之间没有固有的间隔符(空格),需要进行分词处理。引入分词主要是为后继处理做准备,在具体的应用中,要根据具体的情况来选择不同的分词方案。不同分词方案的正确性很大程度上取决于所建的词库,一个词库应具有完备性和完全性两个方面。词库的完备性,简单来说就是对任意一个字串,总能按词库找到对它进行切分的方法;词库的完全性,意味着词库应包含所有的词,建立一个同时满足这两个要求的词库具有很大的难度。而对于某一系统来说,可能只用到其中的一部分。因此在构造词典的时候需要量力而行,在完备和效率之间寻求平衡。

分词后,一般要引入停用词表和高频词表剔除对分类没多大影响的词语。其中停用词指的是那些语法词以及一些虚词、感叹词、连词等;另外有些高频词汇在所有的文本中出现的频率都基本相同,区分性差,也不能作为文本类别的特征。

1.3 特征提取

训练文本和待分类文本经过分词并去除停用词和高频词后,表示文本的向量空间和类别向量的维数也是相当大的,因此需要进行特征项的抽取。特征提取是文本分类系统中十分关键的问题,它可降低向量空间的维数,提高系统的速度和精度,还可以防止过拟合。基于VSM的特征提取方法都是统计的方法,首先利用不同的方法对特征项进行评分。对于待分类文本来说就是计算权重,通过一定的方法计算出权重然后选出分值较高的作为特征构成文本的向量空间。而对于类别矢量空间,常用的特征提取方法有:互信息、信息增益、期望交叉熵和文本证据权等等,文中采用的是互信息特征提取方法。互信息是统计学和信息论中一个重要的概念,它表现了两个统计量间相互关联的程度,关联程度越高,互信息越大,反之亦然。特征项与类别的互信息量可以用下面的公式[5]计算:

$$M(W,C_j)=\log_2\left(\frac{P(W/C_j)}{P(W)}\right)=\log_2\frac{P(W,C_j)}{P(W)\times P(C_j)} \quad (2)$$

其中$P(W,C_j)$是训练语料中特征项W出现在类别C_j中的频率,$P(W)$是训练语料中特征项W出现的频率。为了避免特征项过多造成系统的过拟合现象,计算出所有特征项的互信息量后,要将互信息量从大到小排序,然后选出分值较高的前K个作为特征构成特征向量空间。关于K值的确定需要在实验中进行动态调整。

1.4 训练方法和分类算法

Web文本分类是一个典型的有教师的机器学习问题,一般的可分为训练和分类两个阶段。其中训练算法的工作是对训练文档集合中每篇文本对应的词表进行统计,计算出类别向量矩阵同时进行归一化,最后保存训练得到的向量表,即得到了分类知识库;分类算法(也可称为识别算法)则依据训练得到的分类知识库并用一定的算法对待分类文本进行分类。

文中的训练和分类算法的详细过程如下:

(1)训练算法。

Step 1:对训练文本集合$S\{s_1,\cdots,s_i,\cdots,s_n\}$中的每一篇文本进行分词特征提取得到词语序列$(w_1,w_2,w_3,\cdots)$。

Step 2:根据上文中提到的特征提取方法中的互信息量计算公式计算互信息量。

Step 3:利用多次实验的比较确定提取特征项的数目K,当然对于不同的类别其K值可以不同,最后得到类别向量矩阵同时进行归一化。

(2)分类算法。

Step 1:对于待分类文本集合$T\{d_1,\cdots,d_i,\cdots,d_r\}$中的每一个待分类文档$d_k$,计算其特征矢量$V(d_k)$与每一个$V(c_j)$之间的相似度$\mathrm{sim}(d_k,c_j)$。通常是考虑两个特征矢量之间夹角的余弦,即

$$\mathrm{sim}(d_k,c_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}},$$

其中,W_{ik},W_{jk}分别表示文本d_k和c_j第K个特征项的权值,$1\leqslant k\leqslant N$。

Step 2:选取相似度最大的一个类别作为d_i的类别。

2 扩展的基于向量空间模型的Web文本分类算法

经过上述的训练和分类阶段之后,很显然就可以得到分类结果,并且也可以满足一定用户的需求。但是经研究发现,目前的Web文本分类系统都是将训练方法和分类方法作为分类系统的核心部分,其训练过程就已经决定了系统所具有的分类能力,这种能力在随后的分类过程中是固定不变的,也因此分类后的结果不能达到较高的分类准确率和查准率,不具备不断学习的能力。

鉴于以上存在的问题,文中提出了一种扩展的基于向量空间模型的Web文本分类模型(见图1)和算法,其基本思想是:在传统的训练和分类算法的基础上,加入了一个反馈判定的算法,通过反馈信息将一部分已经得到的分类结果返回,经过修正文本分类器的处理后,调整文本类别矢量,然后再利用相似度测度函数就可以得到新一轮的分类结果。这种方式更加贴近真正意义上的机器学习,使得

该算法具有一定程度上的认知自主性和不断学习的能力。

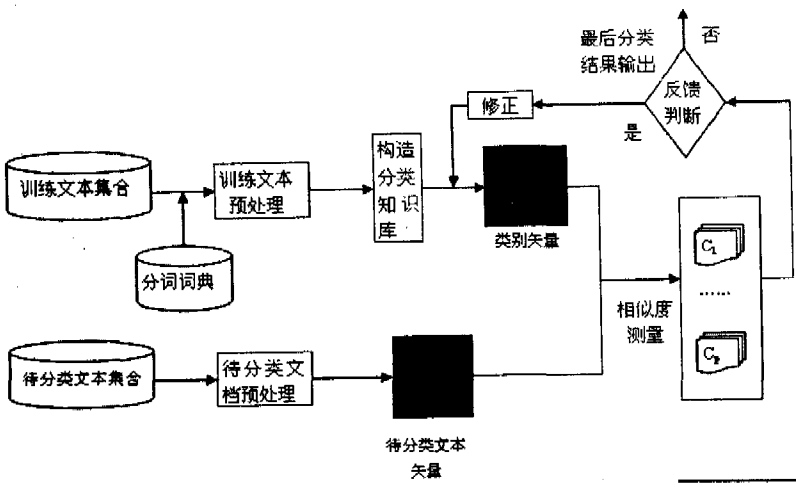

图 1 扩展后的 Web 文本分类模型

该反馈算法的主要过程如下：

Step 1: if $\beta_{TC_i} \geqslant \epsilon_i$ go to Step 2

else go to Step 5

其中 β_{TC_i} 是新文本 T 与对应类别 C_i 之间的相似系数，ϵ_i 是反馈阈值。

Step 2：查询并存储该文本对应的类别 C_i 的类别中心矢量 $C_i(p_{i1}, p_{i2}, \cdots, p_{ik})$和该类别包含的所有特征项数目 K。

Step 3：求出类别 C_i 几何平均前的类别中心矢量 C_i $(q_{i1}, q_{i2}, \cdots, q_{ik})$，其中 $q_{ij} = n \times p_{ij}(j = 1,2,\cdots,k)$。

Step 4：计算类别 C_i 调整后的矢量 $(p'_{i1}, p'_{i2}, \cdots, p'_{ik})$，用新的矢量代替 $C_i(p_{i1}, p_{i2}, \cdots, p_{ik})$，并用合适的数据结构存储起来，其中 $p'_{ij} = \frac{q_{ij} + w_{ij}}{n+1}$

Step 5：过程结束。

3 实验分析

为了验证该扩展分类算法的有效性，从搜虎网站上选取了 300 篇文档，因为搜虎网站已经对这些文档进行了分类，所以可得到大量的训练语料和测试语料。实验主要是在网络游戏、娱乐新闻、体育健身、国际政治、财经 5 个类别之间进行的。将其中的 240 篇文档构成训练文档集合，另外的 60 篇作为测试文本集合。在实验过程中，用到了中科院计算所研制出的汉语词法分析系统 ICTCLAS 进行文本的预处理工作，在此对他们的工作表示感谢。

在对 Web 文档集合进行分类后，得到了第一轮分类结果数据；接着在第一步实验的基础上，利用文中提出的反馈算法，对第一轮的实验结果进行了新一轮的反馈判定过程，在得到反馈信息的同时修改了分类器，即调整了相应的类别矢量，从而得到了新的分类结果。将两个阶段的实验数据进行了比较，得到如表 1 所示的结果数据。

衡量文本分类效果的指标是查全率(recall)和准确率(precision)，其中查全率是被判定为相关的相关文本占全部相关文本的比率；准确率是被判定为相关的文本中真正相关的文本所占的比率。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，即 F1 测试值，其数学公式如下：

$$\text{F1 测试值} = \frac{\text{准确率} \times \text{查全率} \times 2}{\text{准确率} + \text{查全率}}$$

从表 1 可以看出，与第一轮实验结果比较，进行新一轮的反馈判定之后 Web 文本分类的准确率和查全率明显得到提高。

表 1 实验结果数据表

	类别	网络游戏	娱乐新闻	体育健身	国际政治	财经
第一轮分类	准确率(P)	0.915	0.926	0.918	0.746	0.917
	查全率(R)	0.870	0.842	0.963	0.756	0.885
	F1 测试值	0.892	0.882	0.940	0.751	0.901
第一轮分类+反馈算法	准确率(P)	0.923	0.936	0.925	0.755	0.931
	查全率(R)	0.895	0.851	0.970	0.778	0.912
	F1 测试值	0.861	0.891	0.947	0.766	0.921

4 结束语

主要讨论了 Web 文本分类过程中的几个关键技术：文本的表示、词典的构造与分词、特征提取、训练算法和分类算法，并提出了一种扩展的基于向量空间模型的 Web 文本分类模型和算法。通过实验，验证了该种算法在分类在分类准确率和查全率方面的优越性，不过同时也发现该算法在时间效率上是不尽人意的。进一步的工作是研究如何在不提高该算法时间复杂度的前提下来提高分类精度。

参考文献：

[1] Chute C G. An example based mapping method for text categorization and retrieval[J]. ACM Transactions on Information System, 1994,12(3):252-277.

[2] Pawlak Z. Rough Sets[J]. International Journal of Computer and Information Science,1982,11(5): 341-356.

[3] 何 明,冯博琴,傅向华. 基于 Rough 集潜在语义索引的 Web 文档分类[J]. 计算机工程,2004,30(13):3-5.

[4] 张东礼,汪东升,郑纬民. 基于 VSM 的中文文本分类系统的设计与实现[J]. 清华大学学报(自然科学版),2003,43(9):1289-1291.

[5] 陈治纲,何丕廉,孙越恒,等. 基于向量空间模型的文本分类方法的研究与实现[J]. 计算机应用,2004,24(6):277-279.