维普资讯 http://www.cqvip.com

中 文 信 息 学 报
第19卷 第1期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.19 No.1

文章编号:1003-0077(2005)01-0036-06

# 基于向量空间模型的文本分类系统的研究与实现

陈治纲,何丕廉,孙越恒,郑小慎

(天津大学 电子信息工程学院,天津 300072)

**摘要**:文本分类是信息处理的一个重要的研究课题,它可以有效的解决信息杂乱的现象并有助于定位所需的信息。本文综合考虑了频度、分散度和集中度等几项测试指标,提出了一种新的特征抽取算法,克服了传统的从单一或片面的测试指标进行特征抽取所造成的特征"过度拟合"问题,并基于此实现了二级分类模式的文本分类系统。和类中心分类法相比,实验结果表明二级分类模式具有较高的精度和召回率。

**关键词**:计算机应用;中文信息处理;文本分类;测试指标;特征抽取;二级分类模式

**中图分类号**:TP391 **文献标识码**:A

## Research and Implementation of Text Classification System Based on VSP

CHEN Zhi-gang, HE Pi-lian, SUN Yue-heng, ZHENG Xiao-shen

(School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China)

**Abstract**: Text classification is an important research task of natural language processing, which can efficiently resolve the issue of information chaos and help to locate the required information. The traditional approaches of text classification commonly extract feature terms from a single test criterion, which will lead to the problem of "over fitting". This paper comprehensively takes test criterions such as frequency, distribution and concentration into account and proposes a new arithmetic of feature extraction and implements text classification system with two-level mode. The experimental results show that two-level classification mode has higher classification precision and recall compared with center classification method.

**Key words**: computer application; Chinese information processing; text classification; test criterion; feature extraction; two-level classification mode

## 1 引言

Internet 已被公认为是20世纪末人类科技史上的里程碑,它促使人类社会步入了以网络为中心的信息时代。随着 WEB 信息量爆炸性增长,人们很难从大量的信息中迅速有效地提取出所需信息,出现所谓的"信息迷向"的现象。为了准确地定位所需的信息,文本分类的研究显得越来越重要了[1]。现在文本分类在搜索引擎、WEB 页分类管理、电子邮件自动分类、信息过滤等方面都极具应用价值。

本文主要探讨了文本分类系统的关键理论技术和实现。本文组织如下:第一部分为引言,第二部分详细介绍了文本分类的定义及其关键技术问题,第三部分提出了我们实现的基于向量空间模型的文本分类系统的结构框架,第四部分给出了对该系统进行的测试数据和实验设

**收稿日期**:2004-05-21

**基金项目**:天津市科技发展计划项目(023100511)

**作者简介**:陈治纲(1979—),男,硕士生,主要研究方向为信息检索.

维普资讯 http://www.cqvip.com

置,第五部分是实验结果和分析,第六部分是结束语。

## 2 文本分类的定义及关键技术

### 2.1 文本分类的定义

简单地说,文本分类系统的任务是:在给定的分类体系下,根据文本的内容或属性,将大量的文本归到一个或多个类别中。从数学角度来看,文本分类是一个映射的过程,它将未标明类别的文本映射到已有的类别中,用数学公式表示如下:

$f: A \rightarrow B$　　其中,$A$ 为待分类的文本集合,$B$ 为分类体系中的类别集合。

文本分类是系统根据训练集的样本数据信息总结分类规律并确定待分类文本的相关类别。

### 2.2 文本的向量化表示

在自然语言处理领域,文本的表示主要采用向量空间模型(VSM)[2],其出发点是:每篇文本都包含一些用概念词表达的揭示其内容的独立属性,而每个属性都可以看成是概念空间的一个维数,这些独立属性称为文本特征项(常见的特征项类型有字、字串、词、短语等,现有的研究认为以词为单位来进行处理比较合理[3]),则文本就可以表示为这些特征项的集合,这样就不用考虑文本结构中段落、句子及词语之间的复杂关系。因此文本就可以表示成形如 $d = (t_1, w_1; t_2, w_2; \cdots\cdots t_n, w_n)$,其中 $t_i$ 为特征项,$w_i$ 为其对应的权重。特征项权重 $w_i$ 通常用 $tf_{ij} \times idf_i$ 来衡量,定义如下:

$$w_i = tf_{ij} \times idf_i = \frac{tf_{ij} \times \log(N/n_i + 0.01)}{\sqrt{\sum_{t \in d} [tf_{ij} \times \log(N/n_i + 0.01)]^2}} \tag{1}$$

这里 $tf_{ij}$ 为特征项 $t_i$ 在文本 $d_j$ 中出现的频率,$N$ 为文本总数,$n_i$ 为文本集中出现 $t_i$ 的文本数,分母为归一化因子。可见,权重值大的特征项是那些在文本中出现频率足够高,但在整个文本集的其他文本中出现频率足够少的词语,对区别文本具有重要的意义。

### 2.3 特征抽取与选择

特征抽取一般是通过构造一个特征评分函数,把测量空间的数据投影到特征空间,得到在特征空间的值,然后根据特征空间中的值对每个特征进行评估,它可以看作是从测量空间到特征空间的一种映射或变换。特征选择就是根据特征评估结果从中选出最优的且最有代表性的特征子集作为该类的类别特征。因此,特征提取与选择是训练集中文本共性与规则的归纳过程,是文本分类中最关键的问题,它可以降低特征空间的维数,从而达到降低计算复杂度和提高分类准确率的目的。

常用的特征评分函数有:互信息、信息增益、期望交叉熵和文本证据权等等,大量的研究表明采用互信息算法效果明显优于其它算法[4,5]。互信息是统计学和信息论中一个重要的概念,它表征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的公式如下:

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

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。经过比较之后,我们选择互信息量大的特征项作为该类的类别特征。

### 2.4 文本分类的方法

一个好的文本分类方法能够和特征抽取算法相得益彰,取得满意的分类效果。基于向量

维普资讯 http://www.cqvip.com

空间模型的文本分类方法有类中心分类法,贝叶斯算法、KNN 算法和神经网络算法[6]等。其中类中心分类法应用地比较广泛,即在向量空间模型中,文本和类别都被表示为空间中的一个点向量,文本向量和类别向量之间就存在空间上的距离远近,而这种距离就可以采用向量间夹角的余弦来度量,定义如下:

$$SC(d,c)=\frac{\sum_{i=1}^{n}(\omega_{d_i}\times\omega_{c_i})}{\left[\sum_{i=1}^{n}(\omega_{d_i}{}^2)\times\sum_{i=1}^{n}(\omega_{c_i}{}^2)\right]^{\frac{1}{2}}} \tag{3}$$

其中 $d=(d_1,d_2,\cdots,d_n)$为文本 $d$ 的特征向量,$c=(c_1,c_2,\cdots,c_n)$为类别特征向量,即用两个向量之间的夹角的余弦来表示文本与类别之间的相似度,夹角越小,距离越近,余弦越大,相似度越大,反之相似度越小。计算出文本与所有类别的相似度后,将其归入相似度值最大的类别中。本文中我们将以此方法的分类结果为基准,进行研究其它的文本分类方法。

## 3 基于向量空间模型的文本分类系统的实现

### 3.1 文本分类系统的结构框架

从上面的分析中,我们可以知道文本分类过程就是一个建立从文本属性到文本类别空间的映射过程,它主要分为训练过程和分类过程两个阶段。我们实现的文本分类系统的结构框架如下图所示:

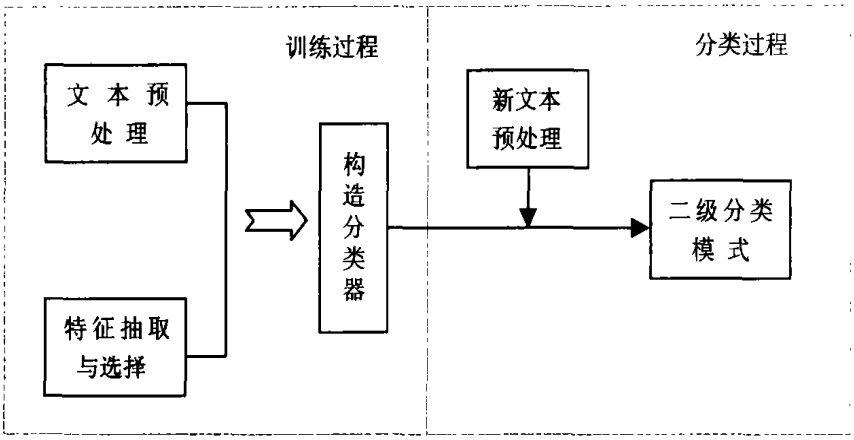


图 1 文本分类系统结构框图

### 3.2 改进的文本特征抽取算法

特征抽取算法的优劣直接影响到文本分类的效果,特征项选择依赖于频度、分散度和集中度等多项测试指标[7]。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;集中度指标认为,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;分散度指标认为,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,而在该类别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。

显然对于某一特征项,其频度越高、分散度越大、集中度越强,则对文本分类越有用,即分辨度越强。从前面互信息法特征评价公式中可以看出,该公式是从频度指标的角度出发,计算每个特征项在每个类别中的出现频度与它在整个文本集中的出现频度的比率,作为该特征项

维普资讯 http://www.cqvip.com

对每个类别分类依据的贡献。这种方法忽略了特征项的分散度和集中度测试指标,从而造成单个特征"过度拟合"的问题。

我们提出了一种新的特征抽取 *FE* 算法,在互信息的特征抽取方法的基础上,给出分散度和集中度测试指标的修正,公式如下:

$$FE(W,C_j)=\log_2\left(\frac{P(W/C_j)}{P(W)}\right)\times\exp(N_j/N)=\log_2\left(\frac{P(W,C_j)}{P(W)\times P(C_j)}\right)\times\exp(N_j/N) \quad (4)$$

其中 $N_j$ 为训练语料中特征项 $W$ 出现在类别 $C_j$ 中的文本数,$N$ 是训练语料中特征项 $W$ 出现的文本数,$P(W,C_j)$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。这样抽取出的特征能很好的体现频度、分散度和集中度测试指标,使其在这些指标中达到整体最优。

### 3.3 二级分类模式

类中心分类法简单直观,但对于类别界限不明显时,该方法性能不高。KNN 算法的效果比较好,该算法的基本思路是:在给定新文本后,选定在训练集中与该新文本距离最近(最相似)的 $K$ 篇文本,根据这 $K$ 篇文本所属的类别判定新文本所属的类别。距离判别一般也采用向量间夹角的余弦来度量,具体定义如公式(3)。如果有多个文本同属于一个类,则该类的权重为这些相似度之和。在新文本的 $K$ 个邻居中,依次计算每类的权重,计算公式如下:

$$p(d,C_j)=\sum_{d_i\in K}SC(d,d_i)y(d_i,C_j) \quad (5)$$

其中 $d$ 为新文本的特征向量,$SC(d,d_i)$为相似度计算公式,而 $y(d_i,C_j)$为类别属性函数,即如果 $d_i$ 属于类 $C_j$,那么函数值为 1,否则为 0。最后比较类的权重并进行排序,将文本分到权重最大的那个类别中。这里 $K$ 值的确定目前没有很好的方法,一般采用先定一个初始值,然后根据实验测试的结果调整 $K$ 值。在本系统中我们结合这两种分类方法,形成了二级分类模式,详细算法如下:

1)对待分类文本进行预处理,包括分词、滤除停用词和文本向量化处理;

2)采用类中心分类法对新文本进行粗分类,依次计算该文本与各类别的相似度;

3)若相似度结果排序的前几位相差较大,则将其归入相似度值最高的类别中;

4)若类别相似度值很接近,满足一定的范围条件时,则在这几个相近类别的训练集中采用 KNN 算法来进行细分类。

## 4 实验设置

### 4.1 测试集

在实验中,我们使用 NTCIR - 2* 的中文资讯检索测试集第一版(CIRB010)作为测试集。该测试集分为三部分:文本集(CIRB010DocumentSet)、查询主题(CIRB010TopicSet)、相关判断(CIRB010RelevanceJudgment ),是由中国时报,中时晚报,工商时报,中央日报,中华日报等五家报纸的新闻资料构成。我们仅使用了中国时报的部分文章来进行测试,这些新闻分为六类:财经(eco)、国际(int)、政治(pol)、社会(soc)、体育(spo)及娱乐(ent)。我们从这六类各抽取 1000 篇文本,其中 500 篇进行文本训练,500 篇是测试文本。

---

* NTCIR 是由日本国家技术委员会组织的国际文本检索评测大会,类似于 TREC,它主要集中在文本检索,Q&A 和文本摘要。一年举行一次,至今已举行了四次。

维普资讯 http://www.cqvip.com

### 4.2 评价方法

文本分类的评价标准类似于信息检索的评价标准，包括精度(查准率)、召回率(查全率)和 $F1$ 测试值：精度是采用文本分类方法分类的正确文本数与实际分类的文本数的比率，即精度 $(precision)=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}$；召回率是采用文本分类方法分类的正确文本数与分类应有的所用文本的比率，即召回率 $(recall)=\frac{\text{分类的正确文本数}}{\text{应有所有文本数}}$；$F1$ 测试值综合考虑了准确率和查会率这两个不同的方面，其公式如下：$F1$ 测试值 $=\frac{\text{精度}\times\text{召回率}\times 2}{\text{精度}+\text{召回率}}$。

## 5 实验结果及分析

我们对每类的500篇文本训练，分别采用互信息和 $FE$ 算法进行抽取与选择类别特征向量，结合类中心分类法对各类的500篇测试文本进行分类比较，实验的结果如表1所示：

**表1 特征抽取与选择实验结果**

| 特征抽取与选择 | | 财 经 | 国 际 | 政 治 | 社 会 | 体 育 | 娱 乐 |
|---|---|---|---|---|---|---|---|
| 互信息法 | 精度 | 0.766 | 0.711 | 0.685 | 0.732 | 0.925 | 0.950 |
| | 召回率 | 0.895 | 0.688 | 0.756 | 0.648 | 0.976 | 0.847 |
| FE算法 | 精度 | 0.784 | 0.765 | 0.713 | 0.746 | 0.931 | 0.954 |
| | 召回率 | 0.902 | 0.703 | 0.796 | 0.658 | 0.980 | 0.862 |

由上表看出应用FE算法应用进行特征抽取，分类精度和召回率均比传统的互信息法有所提高。因此，FE算法能更好的反映文本的分布规律，抽取更有代表性的文本特征项。

下面我们分别采用类中心分类法和二级分类模式对各类的500篇文本进行分类测试，表2为这些文本的分类实验结果。

**表2 文本分类的实验结果**

| 分类法 | | 财 经 | 国 际 | 政 治 | 社 会 | 体 育 | 娱 乐 |
|---|---|---|---|---|---|---|---|
| 类中心分类法 | 精度 | 0.784 | 0.765 | 0.713 | 0.746 | 0.931 | 0.954 |
| | 召回率 | 0.902 | 0.703 | 0.796 | 0.658 | 0.980 | 0.862 |
| | F1 测试值 | 0.839 | 0.733 | 0.752 | 0.699 | 0.955 | 0.916 |
| 二级分类模式 | 精度 | 0.822 | 0.837 | 0.807 | 0.853 | 0.931 | 0.954 |
| | 召回率 | 0.896 | 0.776 | 0.813 | 0.866 | 0.980 | 0.862 |
| | F1 测试值 | 0.857 | 0.805 | 0.810 | 0.859 | 0.955 | 0.916 |

从表中我们可以看出，类中心分类法处理内容主题性较强的类(如体育新闻和娱乐新闻类)，具有很好的分类效果；而处理某些界限不明显的类别(如国际新闻和政治新闻类)时，分类的精度和召回率较低。二级分类模式能够解决这类问题，具有较高的精度和召回率，在分类效果上是最佳的。

## 6 结束语

特征抽取与选择和文本分类方法是文本分类系统的核心，本文综合考虑了频度、分散度和集中度等几项测试指标，提出了一种新的特征抽取算法，克服了传统的从单一或片面的测试指标进行特征抽取所造成的特征“过度拟合”问题，并基于此实现了二级分类模式的文本分类系统。实验结果表明二级分类模式比类中心分类法在精度和召回率方面有很大的提高。今后的工作重点是在本系统的基础上，更深入的结合机器学习、自然语言处理等理论知识，尝试其它

维普资讯 http://www.cqvip.com

分类算法,进一步提高分类效率和精度。

**参 考 文 献:**

[1] 李晓黎,刘继敏,史忠植.概念推理网及其在文本分类中的应用[J]. 计算机研究与发展,2000,37(9):1032-1038.

[2] 刘少辉,董明楷,等. 一种基于向量空间模型的多层次文本分类方法[J]. 中文信息学报,2001,16(3):8-26.

[3] Jian-yun Nie, Jianfeng Gao etc. On the Use of Words and N-grams for Chinese Information Retrieval[A]. Fifth International Workshop on Information Retrieval with Asian Languages[C]. Hong Kong, September 30 - October 1, 2000.

[4] 黄萱菁,吴立德,等. 独立于语种的文本分类方法[J]. 中文信息学报,2000,14(6):1-7.

[5] 秦进,陈芙蓉,等. 文本分类中的特征抽取[J]. 计算机应用,2003,23(2):45-46.

[6] Yiming Yang, An evaluation of statistical approaches to text categorization[J]. In: Journal of Information Retrieval, 1999,1(2):67-88.

[7] 杨允信. 文本文件自动分类之研究[A]. 台湾地区第六届计算语言学研讨会论文集[C],1993.

(上接第7页)

**参 考 文 献:**

[1] 穗志方,俞士汶.汉语单句谓语中心词识别知识的获取及应用[J].北京大学学报(自然科学版),1998,34(2-3):221-229.

[2] 穗志方,俞士汶.面向 EBMT 的汉语单句谓语中心词识别研究[J].中文信息学报,1998,12(4):39-46.

[3] 龚小谨,罗振声,骆卫华.汉语句子谓语中心词的自动识别[J].中文信息学报,2003,17(2):7-13.

**附表:现代汉语词语词类代码表**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 普通名词 | N | 普通动词 | v | 副　词 | d | 时间名词 | Nt | 机构专名 | ni |
| 方位名词 | Nd | 趋向动词 | vd | 代　词 | r | 判断动词 | vl | 数　词 | m |
| 人　名 | nh | 形 容 词 | a | 介　词 | p | 助　词 | u | 拟 声 词 | o |
| 处所名词 | nl | 能源动词 | vu | 连　词 | c | 后接成分 | k | 缩 略 语 | j |
| 地　名 | ns | 区 别 词 | f | 叹　词 | e | 前接成分 | h | 字 符 串 | ws |
| 普通专名 | nz | 量　词 | q | 习 用 语 | i | 标　点 | w | | |