·工作研究·

基于词频的中文文本分类研究

姚兴山
(南京大学信息管理系，江苏 南京 210093)

〔摘 要〕本文对中文文本分类系统的设计和实现进行了阐述，对分类系统的系统结构、特征提取、训练算法、分类算法等进行了详细的介绍。将基于词频统计的方法应用于文本分类。并提出了一种基于汉语中单字词及二字词统计特性的中文文本分类方法，在无词表的情况下，通过统计构造单字和二字词表，对文本进行分类，并取得不错的效果。

〔关键词〕词频统计；特征选取；中文文本分类
〔中图分类号〕TP393 〔文献标识码〕A 〔文章编号〕1008-0821(2009)02-0179-03

Chinese Text Classification Based on Word Frequency Statistics

Yao Xingshan
(Department of Information Management, Nanjing University, Nanjing 210093, China)

〔Abstract〕In this paper, the designation and accomplishment of a Chinese text classification system was described, and system construction, feature selection, training arithmetic, classification arithmetic were introduced. The methods based on word frequency statistics were used in Chinese text classification. At the same time, a new Chinese text classification method was introduced in this paper, which based on word and two-word statistical properties. In the absence of vocabulary, statistics through word structure and the second word list, text classification, and achieved good results.

〔Key words〕word frequency statistics; feature selection; chinese text classification

在中文信息处理中，中文文本分类（Text Categorization，简记为TC）问题一直是重要的研究内容。中文文本分类的最终目标：在给定的分类体系下，根据训练文档集合，自动确定新的文档的类别。通过文档的自动分类，实现资源从无序到有序的组合，以便用户高效的利用资源[1]。早期的自动文本分类以知识工程的方法为主，根据领域专家对给定文本集合的分类经验，人工提取出一组逻辑规则，作为计算机自动文本分类的依据。本文在论述中文文本分类系统实现过程中，避免了过去那种机械分词的步骤，在文本特征抽取的过程中引入统计学的方法，提出了一种基于单字和二字统计特性的中文文本分类方法，在文中详细的介绍了流程和机器学习的过程，最后给出了实验测试的结果。

1 系统设计

从数学的角度可以把文本分类看成是一个映射的过程。它将未标明类别的文本映射到已有的文本中，该映射可以是一一映射，也可以是一对多映射，一篇文档和多篇文档相关联。用数学的公式表示为：$F(A)=>(B)$其中 A 为待分类的文本集，B 为分类体系中的类别集，F 为文本分类规则。图1我们给出了中文文本分类的流程图。

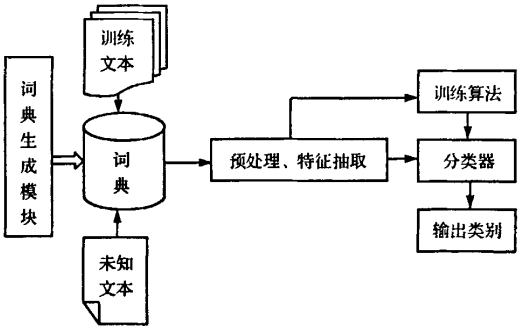

图1 中文文本分类系统流程图

收稿日期：2008-10-14
作者简介：姚兴山（1974-），男，博士研究生，研究方向：信息检索、数据挖掘及中文信息处理，发表论文5篇。

系统主要由词典生成模块、训练模块和分类模块组成。词典生成模块通过对文本中单字字频信息以及相邻字的共现信息进行统计，产生分词词表。训练模块首先对训练文本进行预处理，然后进行特征选取和参数训练，最后生成文本分类器。分类模块通过对待分类文本的预处理及特征选择后，由文本分类器自动对文本进行分类。

1.1 词典生成模块

我们通过向量空间模型（Vector Space Model）把文本表示成一个由词条向量组成的向量空间，每个文本 d 都可以映射为空间的一个特征向量 $V(d)=(T_1,W_1(d),T_2,W_2(d),\cdots T_n,W_n(d))$，其中 T_i 表示特征项，$W_i(d)$ 表示对应分量的权重[2]。

在中文文本中可以采用字、词或者短语作为表示文本的特征项，由于短语的切分难度要比字和词的切分难度大的多，所以我们采用字和词作为文本的特征项。那么所要做的第一步就是要进行文本的分词，如果把所有的词都作为特征项，那么特征向量的维度将过于巨大，从而导致计算量巨大，要完成文本分类几乎是不可能的。笔者摒弃了过去的那种传统的词典分词方式，引入统计学的模型，通过计算词频来实现特征项的抽取。根据文献［3］的统计见表 1：

表 1 词条分布情况表

词条字数	1	2	3	4	5	6	7
词条数	9 919	65 891	26 352	21 699	5 124	2 446	980
出现频率%	56.75	39.65	2.21	1.19	0.144	0.083	0.023

由表 1 可见，在汉语常用词中单、双字词出现的频率最高，同时数量上也占了绝大部分。

从语言学的角度分析。不同类别的文本中，汉字的分布是有规律的，研究人员大量的实验证明，用单字词统计特性来进行文本分类精度可以达到 65% 左右。这说明采用单字词来表示文本特征进行文本分类精度还是可以保证的，如果再结合少量的对类别区分能力强的二字词的话，则应该可以提高系统的精度。因此我们采用单字词为主结合少量的二字词为文本的特征向量，相应的词表也是由单、双字词构成。系统包括的分类体系有军事、体育、财经、新闻和娱乐 5 个类别，我们采用国内知名搜狐门户网站上的网页文本作为训练文本集。

首先进行单字词的统计，把待统计的所有的文本组合为一个大的文本 A，对文本中的所有汉字出现频率进行统计并排序。由于汉语中的常用汉字大概在3 000个左右，为了确保单字词对文本的覆盖率，只要出现过的汉字就作为单字词保留在单字词表中。接下来进行二字词的统计。文献［4］提出了通过计算相邻汉字的互信息并以此建立词表。这种方法运算过于复杂，当文本量很大时速度也比较慢。由于我们目的是找到各类文本中的常用的二字词，因此我们采用更为简单的方法，通过计算相邻二字共现的频率（文中以 WF 表示）来查找二字词。

1.2 训练模块

在进行中文文本处理前我们需要首先对文本进行预处理。由于我们需要抽取的特征向量都是词，对于英文、数字等一些符号都不予考虑。同时对那些频率极高却没有什么实际意义的词，如："的"、"了"、"和" 等。这些词对文本分类的影响甚小，因此通过停用词表剔除。

文本表示中 T_i 及其 W_i 选取称为特征抽取。特征算法的优劣影响到文本分类的效果。我们在系统中采取了基于词频统计的特征抽取方法。通过大量的文本训练，根据特征项对文本内容的贡献，经过多次的统计学习得出权重评价。以这种方法选择的特征项集具有如下的两个特点：

（1）完全性：特征项能够体现目标内容；

（2）区分性：根据特征项集能够将目标同其他文本相区别。

根据这两个特点可得，当一个词条在某类文本中出现的频率越高。而在其它类别的文本中初相的频率越低，则该词条在该类别特征向量中的权重就越大。因此我们构造了词条权值评价函数：$W_{ik}=tf_{ik}\cdot\log^2(a*N_k/n_k+0.05)$[5]其中 tf_{ik} 表示词条 T_k 在文本 D_i 中出现的频数，N_k 表示本类别训练文本中出现 T_k 的文本数，n_k 表示词条 T_k 在所有的文本中出现的频数，a 为系数可以根据实验的结果进行调整。为增强文本类特征的稳定性，我们取各类别文本的重心作为该类的特征向量，文本类重心定义为一类文本中所有文本向量的平均向量，第 k 个类的重心记为 $C_k=(C_{k1},C_{k2},C_{k3},\cdots\cdots C_{kn})$，$n$ 为向量空间的维数，m 为类 k 中文本的数目，则有公式 1：

$$C_{kj}=\sum_{i=1}^{m}W_{ij}/m \tag{1}$$

W_{ij} 表示文本 D_i 的第 j 个项的权重。在实际的使用中，为了降低个别高频特征项对其它中低频项的抑制作用，我们对特征向量进行了归一化处理。

1.3 分类模块

根据各文本的中心特征向量对未知文本进行分类的技术关键是分类算法。我们系统采用了向量最小距离法计算待分类文本与各类别的相似度，并把该文本归入相似度最大的类别。其过程就是通过计算未知类别文本的特征向量和各类别的中心特征向量之间的夹角余弦。其计算公式见公式 2[6]：

$$Sim(V,U)=COS(V,U)=V\cdot U/|V||U|=\left(\sum_{k=1}^{n}W_{uk}W_{uk}\right)\Big/\sqrt{\left(\sum_{k=1}^{n}W_{uk}^2\right)\left(\sum_{k=1}^{n}W_{uk}^2\right)} \tag{2}$$

其中 W_{ik}表示文本向量 V 的第 k 个特征项的权值，W_{uk}表示文本向量 U 的第 k 个特征项的权值，$V \cdot U$ 表示向量 V 和 U 的点积。分类的过程如图 2 所示。

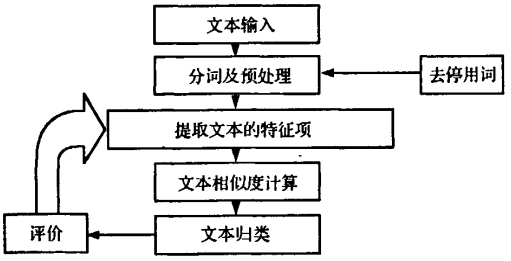

图 2

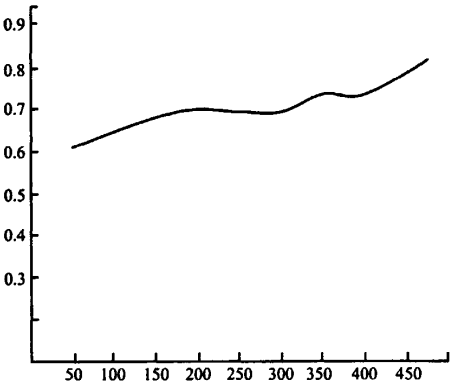

图 3 文本量和分类准确率拟合曲线

2 数据测试及讨论

由于没有标准的测试文本集，我们采用一些大的门户网站的各类别的网页作为我们的测试数据。我们从搜狐网站上下载了 5 种类型的网页：娱乐、财经、体育、军事和新闻。这些网页都是已经经过网站分类好的放在对应的类别中。每种类别下载 500 篇文本，按照 4:1 的比例分为训练文本和测试文本。我们进行了 2 种类型的实验，一种是以单字为文本的特征，另一种是以单字加二字词为文本的特征，测试的结果如表 2 所示。

表 2 文本分类测试结果

测试类别	单字词精度	单字 + 二字词精度
娱乐	68%	82%
财经	70%	80%
体育	76%	85%
军事	78%	84%
新闻	75%	85%

我们同时也做了一个这样的测试，训练文本集的大小对分类精度的影响是如何变化的，我们用曲线进行表示，用 100 篇文本作为基点，用来作为测试的文本是统一的，其变化的曲线如图 3 所示。

由于我们实验的规模小，用来进行训练和测试的样本数量不是很大，实验的数据可能存在一定的误差。但是我们从实验的结果可以看出，利用单字词和二字词的结合其文本分类的精度有了显著的提高。

3 结束语

本文中我们提出了一种中文文本分类系统的设计和实现的过程，给出了分类的流程图，引入统计词频的方法，实现了以单字结合二字的中文文本分类的方法，这样的系统对于分类的精度要求不是很高的需求是完全满足的。由于在文本的特征选取的过程中并没有结合其它的特征选取方法，并且在统计的过程中对低频词的处理不是很好，导致我们的特征的选取未必是最准确的，文本分类的精度受到了影响。在接下来的研究当中我们将重点关注信息增益 IG[7]、互信息 MI、x^2 分布（CHI）、期望交叉熵、优势率、文本证据权等这些特征选取法[8]，根据其不同的特性进行组合，以实现其优势的互补。尝试的引进反馈和层次分类机制，进行阈值的调整研究，改进分类算法，以实现文本分类精度的提高。

参 考 文 献

[1] 郝晓燕，常晓明. 中文文本分类研究 [J]. 太原理工大学学报，2006，36 (6)：710－713.

[2] 金凯民，苗夺谦，段其国，等. 一种基于隐含子类信息的粗糙集中文文本分类方法 [J]，计算机科学，2008，35 (2)：147－149.

[3] 李庆虎，陈玉健. 一种中文分词词典新机制一双字哈希机制 [J]. 中文信息学报，2003，17 (4)：13－18.

[4] 黄晓红，康松林，朱晓娟，等. 基于词频统计的中文分词研究 [J]. 计算机工程与应用，2005，(7)：67.

[5] 熊忠阳，黎刚，陈晓莉，等. 文本分类中词语权重计算方法的改进与应用 [J]. 计算机工程与应用，2008，44 (5)：187－189.

[6] 刘博，杨柳，袁芳，等. 改进的 KNN 方法及其在中文文本分类中的应用 [J]. 西华大学学报：自然科学版，2008，27 (2)：33－36.

[7] 鹿苏玲，蔡庆生. 中文文本分类中的特征选择研究 [J]. 计算机仿真，2007，24 (3)：289－291.

[8] 胡燕，吴虎子，钟珞，等. 中文文本分类中基于词性的特征提取方法研究 [J]. 武汉理工大学学报，2007，29 (4)：132－135.