基于统计分词的中文文本分类系统

①吴雅娟　柳培林　②丁子睿

①大庆石油学院 计算机与信息技术学院,黑龙江 大庆　163318

②大庆石油管理局 通信公司,黑龙江 大庆　163453

摘要:本文阐述了一个中文文本分类系统的设计和实现,对文本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介绍,将基于统计的二元分词方法应用于中文文本分类,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,实现了在事先没有词表的情况下,通过统计构造单字及二字词词表,从而对文本进行分词,然后再进行文本的分类。

关键词: 中文文本分类 统计分词 特征提取

Chinese Text Classification System Based On Statistical Word Segmentation

WU Ya-juan　LIU Pei-lin　DING Zi-rui

①Computer Science and Engineering College, Daqing Petroleum Institute, Daqing, Heilongjiang

②Communication Company, Daqing Petroleum Conservancy, Daqing, Heilongjiang

Abstract:In the article I described the designation and accomplishment of a Chinese text classifycation system,and introduced system construction? feature selection? training arithmetic and classification arithmetic,achieved the goal that comminute word on condition that having no vocabulary.

Key Words: Chinese Text Classification; Statistical Word Segmentation; Feature selection

中图分类号:TP391 **文献标识码**:A

1 前言

文本分类属于人工智能技术和信息获取技术相结合的研究领域，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。进入上世纪九十年代以来,基于统计的自动文本分类方法日益受到重视，它在准确率和稳定性方面具有明显的优势。本文主要论述一个中文文本分类系统的设计和实现技术，在词表的构建过程中提出了基于词频的统计方法，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，还详细介绍了系统流程和机器学习的过程。

2 系统设计

文本分类是指在给定的分类体系下，由计算机自动对已知类别的样本进行学习，并且总结出不同类别文本的特征作为判别依据，然后根据文本的内容特征

他们的学习兴趣,调动他们的学习积极性和主动性。

参考文献

[1]武　兵.印刷色彩[M].北京:中国轻工业出版社,2002.

[2]博嘉科技.中文版 Photoshop 7.0 平面培训教程.北京:中国铁道出版社,2002

[3]钟玉琢 沈洪 吕小星编著.多媒体技术及其应用.北京:机械工业出版社,2003

收稿日期:2005 年 4 月

自动判别文本类别的过程。

从数学角度看，文本分类是一个映射过程，它将未标明类别的文本映射到已有的类别中，该映射可以是一一映射，也可以是一对多的映射，因为通常一篇文本可与多个类别相关联。用数学公式表示为：F (A)=>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，F 为文本分类的判别规则。

图 1 给出本文实现的中文文本分类系统的流程图。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个汉字的字频信息以及相临汉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，然后进行特征选择和参数训练，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。

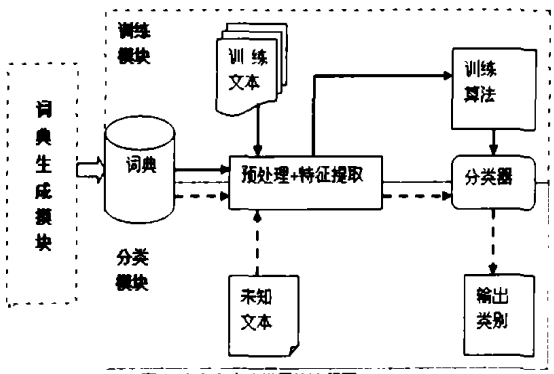

图 1 中文文本分类系统流程图

图 1 中文文本分类系统流程图

2.1 词典生成模块

对中文文本进行处理的第一步是把它表示成计算机可以识别的数据结构，目前较成熟的文本表示方法是向量空间模型(Vector Space Model)，在该模型中，文本空间被看作一组由词条向量张成的向量空间，每个文本 d 都可以映射为此空间的一个特征向量：V (d)=(T_1,W_1(d),T_2,W_2(d),……T_n,W_n(d))，其中 T_i 表示特征项，Wi(d)表示对应分量的权重。

特征项通常是字或者词，由于汉语的书写形式是没有分隔标记的汉字串，词与词之间没有显著的界限，因此，要把中文文本表示成特征向量，就必须对文本进行分词处理。当前自动分词普遍采用的方法是使用一个预定义的通用词表对文本进行切分。

这种方法的主要缺点是为了保证词语的覆盖范围，往往需要一个非常大的词表，这必将大大增加系统的运算复杂度，而且使用固定词表分词无法解决不断出现新词这一问题。

于是有人提出一种基于统计语言模型的分词方法，这种方法计算一些统计量，根据这些数据构建一个适用于和语料库文本同类型的文本的分词词表，从而进行分词。

这种方法能发现一些新词。如在新浪网站众多的文本中，使用统计方法可以发现“新浪”应该被视为一个词，这一点对于分词词表的更新具有特别的意义。由于网上文本内容数量众多，对于一些文本分类任务来说，可以很容易地获取大量同类型文本用于统计和训练。

综合以上几点，本系统应用第二种方法进行分词。

汉语中的词按构成词的汉字的个数不同，可以被分为单字词、二字词、三字词以至 n(n>3)字词。根据现代汉语频率词表[1]的统计，在最常用的 9000 个词中，单字词占 26.7%，二字词占 69.8%，三字词占 2.7%。这个结果显示，在汉语常用词中单字词和二字词占绝大部分。

从语言学角度分析，不同类别的文本中，汉字的分布是有规律的，并且这种规律是相对稳定的，研究人员经过大量的实验证明，只用单字词的统计特性进行文本分类，精度能够达到 80%左右，这说明以单字词作为表示文本的主要特征进行分类是可行的，精确度是有保证的，如果再结合对类别区分能力强的少量二字词，一定能使系统精度有所提高。

因此，本系统采用以单字词为主，附加少量二字词作为表示文本的特征向量，相应的，分词词表也由单字词和二字词构成。

本系统采用的分类体系包括军事、体育、科技、娱乐、财经五个类别，采用网络上大量已知类别的文本作为训练文本。本模块就是对一部分训练文本进行统计以产生分词词表。

首先进行单字词统计，把待统计的各类文本组合成一个大文本 A，对文本中所有汉字的出现频率进行统计，并从大到小进行排序，由于汉语中常用汉字只有 3000 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。然后进行二字词统计，论文[2]提出了一种通过计算两个相邻的汉字之间的互信息，并据此建立二字词表进行分词的方法，该方法运算较复杂，当数据量较大时运算速度非常慢。

由于我们的目的是找到各类别文本中较常用的二字词，因此本文提出一种简单、实用的二字词产生方法，即通过统计相临汉字共同出现频率（文中表示为WF)的方法查找二字词，例如：a、b是前后相临的两个汉字，当统到a时把ab当作二字词，然后在文本A中统计ab的出现次数作为WF值，如果WF大于某个设定值，则认为ab是二字词，把ab保存在二字词表中，如果WF小于设定值则认为ab不是词。

统计完成后形成的二字词表中保留的是文本中出现频率较高的二字组合。虽然有些组合并不是严格意义上的词，但它们对分类的作用是一样的，所以本文把它们也当作词。最后，把单字词表和二字词表连接在一起组成分词词表。

2.2 训练模块

2.2.1 文本预处理

在进行中文文本处理以前，先要对其进行预处理，分词是其中很重要的一个环节，本系统采用机械分词法，依据词典生成模块产生的分词词表信息，以字符串匹配原理进行分词，分词时按照一定的策略将汉字串和分词词表中的词条逐一进行匹配，如果匹配成功就加以切分。

按照对字扫描方向的不同，机械分词方法可分为正向匹配和逆向匹配，按照不同长度词的优先情况，可分为最大匹配和最小匹配，本系统采用正向最大匹配法。有些词在任何类别的文本中出现频率都极高，而且没有实际含义，例如："的"、"了"、"和"、"然后"等等。这些词对于分类几乎没有影响，因此在分词完成后利用停用词表剔除这些无用词。

2.2.2 特征提取

文本表示中词条Ti及其权值Wi的选取称为特征提取。特征提取是文本类共性与规则的归纳过程，也是系统的训练过程，是分类系统的核心，特征提取算法的优劣直接影响到文本分类的效果。在本系统中采用的是基于词频统计的特征提取方法。

权重评价需要在大量训练文本的基础上，根据各特征项对文本内容的贡献，经过多次统计学习完成。自然语言文本中，各词条在不同内容的文本中所呈现出的频率分布是不同的，因此我们可根据词条的频率特性进行权重评价。一个有效的特征项集，必须具备以下两个特征：

- 完全性：特征项能够体现目标内容。
- 区分性：根据特征项集，能将目标同其它文本相区分。

根据以上两个特征可得，词条的重要性正比于词条在本类别文本内出现频数，反比于在所有训练文本内出现该词条的频数，也就是当一个词条在某类文本中出现频率越高，而在其它类别文本中出现频率越低时，则该词条在该类别特征向量中的权值越大。因此我们构造出词条权值评价函数：

$$W_{ik}=tf_{ik} \bullet \log_2(a*N/n_k+0.05) \qquad (1)$$

其中tf_{ik}表示词条T_k在文本D_i中的出现频数，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在所有训练文本中出现的文本频数，a为系数，可根据实验结果进行调整。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$C_k=(C_{k1},C_{k2},\cdots,C_{kn})$，n为向量空间维数，m为类k中文本数目，则有

$$C_{kj}=\sum_{i=1}^{m} W_{ij}/m \qquad (2)$$

W_{ij}表示文本D_i的第j个项的权重。实用中，为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。

2.2.3 分类模块

本模块的功能是根据各文本类别的中心特征向量，对未知文本进行分类。技术关键是分类算法的选择。本系统采用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。向量最小距离法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夹角的余弦，判定文本与类别的相似程度，余弦值越大相似度也越大。相似度计算公式如下：

$$Sim(V,U) = COS(V,U) = V \bullet U / |V||U|$$

$$= (\sum_{k=1}^{n} W_{vk}W_{uk}) / (\sum_{k=1}^{n} (W_{vk})2)0.5 (\sum_{k=1}^{n} (W_{uk})2)0.5$$

其中，W_{vk} 表示文本向量 V 的第 k 个特征项的权值，W_{uk} 表示类别特征向量 U 的第 k 个特征项的权值，$V \bullet U$ 表示向量 V 与 U 的点积。分类过程描述如下：

步骤 1：读入待分类文本并进行分词等预处理。

步骤 2：对文本特征进行筛选，形成该文本的特征向量。

步骤 3：计算该特征向量与各类别的中心特征向量的相似度。

步骤 4：把文本归入相似度大的类别。

3 实验结果

由于没有通用的数据集，我们采用新浪网站的各种类别的新闻网页作为试验用数据集，新浪网站用户访问率高，中文新闻内容丰富，在中文新闻网页中比较具有代表性。我们从新浪网站下载了五类网页，包括：娱乐、财经、体育、军事和科技，这些网页均是由网站工作人员手工进行分类并放置到相应的目录中的。

每一类下载 300 篇文本，按 4:1 分为训练文本和测试文本。我们进行了两种实验，一种是以单字作为文本特征，另一种以单字加二字词作为文本特征，结果如表 1 所示：

以单字为特征的分类精度 以单字加二字词为特征的分类精度

表 1 不同文本特征的分类结果对比

娱乐	78%	82%
财经	75%	80%
体育	87%	90%
军事	82%	85%
科技	80%	87%

从实验结果可以看出，以单字和二字词两种特征表示文本的分类精度，比以单字一种特征表示文本的分类精度有明显提高。

虽然实验的规模较小，并且文本类别比较少，所得出的结果肯定存在一定的偏差，但至少说明二字词的加入确实可以提高分类的精确度。

4 结论

本文主要讨论一个中文文本分类系统的设计和实现过程，提出了分类系统的流程图，并分别介绍了系统的各个模块。引入统计分词的方法解决了没有通用词表的问题。提出一种以单字词为主附加少量二字词为特征的中文文本分类方法。

将来研究工作打算改进特征提取算法；尝试几种分类算法，并比较其性能；尝试引入反馈机制；进行阈值调整研究；引入层次分类机制，进一步改进算法。

参考文献：

[1] 王还，常宝儒. 现代汉语频率词典.北京语言学院出版社,1986

[2] 殷建平. 汉语自动分词方法. 计算机工程与科学,1988,20(3)

[3] 黄萱菁，吴立德. 独立于语种的文本分类方法. 2000 International Conference on Multilingual Information Processing,2000,37-43.

[4] 鲁 松，白 硕，等. 文本中词语权重计算方法的改进. 2000 International Conference on Multilingual Information Processing,2000,31-36.

[5] Yiming Yang. An evaluation of statistical approaches to text categorization. In Journal of Information Retrieval,1999,Vol 1, No.1/2:67--88.

[6] David D.Lewis, Marc Ringuette. A comparison of tow learning algorithms of text categorization. In Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas,NV,1994:81-93.

[7] Andrew McCallum, Kamal Nigam. A comparison of event models for naive bayes text categorization.AAAI-98 Workshop on "Learning for Text Categorization",1998.

[8] Yiming Yang ,Xin Liu. A re-examination of text categorization methods. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) , 1999,42-49.

收稿日期：2005 年 4 月