基于标题类别语义识别的文本分类算法研究

付婷婷 张 磊

（北京交通大学海滨学院计算机与信息技术学院 河北 沧州 061199）

[摘 要] 作为当前热门的信息处理课题之一，文本分类以先进的互联网技术和电子计算机技术为基础，通过机器学习来达到良好的信息定位效果，保证使用者能够在海量的信息中迅速进行筛选有利信息，极大地提升工作效率。目前，文本分类涉及领域不断扩充，逐步在新闻、舆情、邮件和信息检索等领域开始投入使用，并取得了良好的效果。就目前情况而言，其自身的研究意义极强，并且整体的应用前景趋于乐观。

[关键词] 标题类别；语义识别；文本分类；算法；研究

随着信息处理工作的不断优化升级，机器学习技术逐步趋于完善，文本分类工作的便捷性与高效性特征明显。在当前的文本分类过程中，其主要依托量化空间模型技术来进行文档信息读取与筛选工作，通过降低维度来生成向量，并保证整体的词表能够处于可应用状态。同样，在文档的分类与主体整合过程中，主题模型主题逐渐趋于多元，其中的词汇与词句是整体主题组成的核心与关键。通过有效的建构模型，语义信息被突出显示，相应的语义维度由此产生。当前，主题与VSM模式是主要的词袋模型，文本信息相关性不断提升，经过有效的排列，语义信息凸显，文本分类工作结束。

一、文本分类算法描述

类别标签是文本分类的结果，其往往通过对文档进行适当的分类与整合并最终确定标签来实现良好的文本分类效果。作为一种监督性学习行为，文本分类可以通过标注子集来进行函数映射工作，并通过类别预估来不断优化样本行为，从而得到一段完整的映射函数。其中，具体的处理流程如图1所示。在进行文本分类的过程中，测试阶段与训练阶段是其主要运行机制。测试阶段通过文档处理来进行特征选择，在优化分类设备的泛化能力基础上不断提升函数数值的精准性，保证测试效果最优，函数损失最少。训练阶段主要通过优化特征选择来保证分类设备自身的数据量不断提升，在最少损失函数的前提下提升文本分类效果。

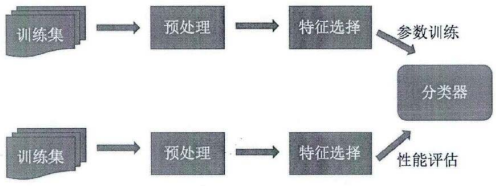

图1 文本分类流程

二、文本标题对文本类别主题的指向分析

从读者角度来看，标题的指向性是文本类别主题的关键，通过区分正文与辩题，整体文档的预示性不断提升。因为正文与文本是文章的主要组成内容，二者的有机结合提升了文章的结构性与完整性，其中，标题扮演文章的统领角色，能够达到有机整合与高度概括的效果。正文内容是文章的解释与重要组成，能够发挥阐释作用。

话题到述题是文章知识结构的拓展模式，语言学家将其归纳为新型篇章语言结构。因此，我们可以将标题作为文章的话题，将正文作为文章的述题。作为叙述的起点，标题能够有效的引导读者阅读，并阐述文章所要体现的价值观与内容。作为叙述的重心，正文是文章的主要叙述内容，可以达到引领与带动阅读的效果，并切实提升读者的理解性。

在文章主题方面，标题与文本一直呈现出相互影响的辩证统一关系，二者共同引导着读者对于文章的阅读与理解，从而发挥出一定的指向作用。其中，可以归纳为下面几类情形。第一，标题引领主题。在文章《反对自由主义》和《鱼雷战》种，读者可以直接通过标题确定文章的主题和叙写内容，其中的主题词椭圆、自由、鱼雷和曲线等字眼同样可以直接进行主题确定。第二，标题引领文章性质。在文章《中国雷达五十一年》和《糖尿病患者运动的好处》中，标题将文章的内容进行明确且细致的确定与限定，标题中的雷达、糖尿病等字眼自身的指向意义极强强烈。第三，标题引领文章的主要线索。在文章《中国经济学百年回顾》和《胶片和前卫电影》中，标题将文内容进行一定的阐释自身的指向性同样较强，读者在读到经济学和电影等字眼时，就可以确定文章的主题意义。

综上所示，标题是文章的关键，是文本内容的核心与重心，大多数文章内容会结合标题进行阐述。就科技论文而言，经过完整的抽样统计我们可以发现科学论文的文章主题99%以上都由标题进行直接反映，其中95%的科学论文标题都能够直接预示文章主题。通过数据分析我们可以发现，文本标题自身的指向性极强，因此，文本分类工作同样需要以文本标题作为基础进行准确判断。

三、文本标题识别算法

为提升文本分类效率，应用类别文本标题中的特殊性语言是优化与提升识别性的重中之重。在语言特色方面，文本标题自身的标记性极强，通过确定位置、形式和词语即可保证信息的识别性大幅提升，文本识别效率同样能够得到不断优化。就标记性的词语而言，其常常出现在正文中，并且自身的结构与句法格式较为特殊，可以通过标点符号来进行文字内容的甄别与筛选。通过上述特征，建立健全完整的识别知识库是提升文本辨别效果的关键。

在进行文本分类算法选择过程中，需要充分结合类别信息自身特有的特征来进行，可以采用高频词、概念词来降低特征集维数，保证分类信息更加完善，通过简化程序来提升信息完整度，从而达到良好的文本分类效果。

四、结束语

随着科学技术的不断提升，文本分类技术逐渐趋于完善，依托先进的TCSR算法，文本分类工作取得了良好的效果。通过充分整合文本标题信息，文本分类精度不断优化升级，自身的拓展性逐渐提升。为此，在未来的文本分类工作中，只有不断拓展与创新，才能更好地进行方法优化与学习整合工作，并逐步形成良好的发展循环。

参考文献

[1]刘英涛.短文本分类研究[D].重庆理工大学，2016.

[2]袁彬.基于语义特征的文本分类算法研究[D].北京邮电大学，2016.

基金项目：沧州市重点研发计划指导项目，基于Bayes信念网络模型的多维向量信息检索的应用研究，183103008

沧州市科学技术研究与发展指导计划项目，大规模社会网络的社区发现与应用研究，172103002

河北省应用技术大学研究会课题，独立院校《数据结构》启发式教学方法的研究与应用，JY2018024