ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.12, No.34, December 2016
E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

文本分类及分类算法研究综述

张磊

（湘潭大学，湖南 湘潭 411105）

摘要：文本分类是一种能在海量文本信息资源中快速准确地获取所需信息的机器学习方法。文中阐述了文本分类的发展进程以及文本分类的一般流程，根据不同的文档特征对文本信息资源进行分类并对几种常见的文本分类算法进行了对比分析。

关键词：文本分类；文本内容；情感倾向；分类算法

中图分类号：TP311 文献标识码：A 文章编号：1009-3044(2016)34-0225-02

DOI:10.14004/j.cnki.ckt.2016.4756

The Research Summary of Text Categorization and Classification Algorithms

ZHANG Lei
(XiangTan University, Xiangtan 411105, China)

Abstract: Text categorization is a machine learning method that can obtain the required information quickly and accurately in the mass text information resources. Text categorization's development of the process and the general process of text categorization are discussed in the paper, text information resources are classified according to different characteristics of documents and several common algorithm of text categorization is analyzed in the paper.

Key words: text categorization; text content; emotional tendency; classification algorithm

1 背景

随着互联网在社会中的大规模应用，网络上的信息资源正在以指数级爆炸增长，Web已经成为一个规模十分庞大的信息资源库。在各种形式的信息中，非结构化的文本信息仍然是十分重要的信息资源之一。在海量的文本信息中，获取最有效的信息资源是信息处理的基础，而文本分类能更好地帮助人们组织管理好海量的文本信息，快速准确地获取所需信息，实现个性化的信息。文本分类在众多领域中均有应用，常见的应用包括：邮件分类、网页分类、文本索引、自动文摘、信息检索、信息推送、数字图书馆以及学习系统等。

2 文本分类

2.1 文本分类的发展进程

文本分类是指根据预先定义的主题类别，按照一定的规则将文档集合中未知类别的文本自动确定一个或几个类别的过程[1]。回顾文本分类的相关研究，可追溯到20世纪60年代，那个时期Maron开创性的提出了概率索引模型，采用了贝叶斯公式来进行文本分类[2]。到了20世纪80年代，主要通过各领域专家提供的知识形成规则，手工建立文本分类器，在这个时期采用的文本自动分类方法主要是基于传统的知识工程。直到20世纪90年代，以机器学习和统计方法为基础的文本分类技术逐步发展起来并不断完善了人工建立分类器的不足，使得文本分类更加准确、有效。目前，文本分类也已经开始应用于对国内中文文本的研究，在Web文档自动分类、自动文摘、数字图书馆等诸多领域开始了应用。

2.2 文本分类的一般流程

文本分类是根据已被标注的训练文本集，通过特征选择、特征提取等方法得到特征项或者根据文本表示方法通过训练得到文本类别间的关系模型即文本分类器，然后用训练得到的关系模型对测试文本集进行文本表示得到文本分类结果的一个有指导的学习过程。文本分类的过程可分为训练和分类两个过程，大致分为文本表示、分类器构建和效果评估三个步骤。文本分类的一般流程如图1所示：

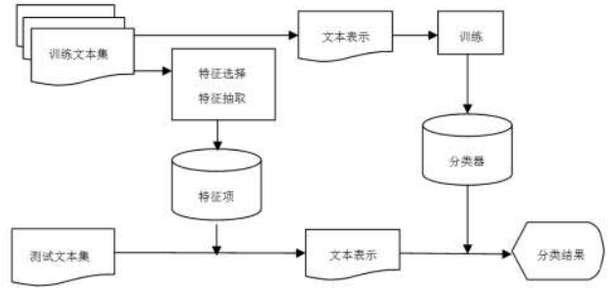

图1 文本分类的一般流程

2.3 文本的分类

在具体的研究中，根据文档的特征不同可以将文本按不同的方式进行分类，如按文本标题、文本内容、情感倾向性、文本风格等方式进行分类。

1）按文本标题分类

收稿日期：2016-10-29

作者简介：张磊（1990—），女，江西上饶人，硕士研究生，主要研究方向为管理信息系统。

按文本标题分类顾名思义即根据文本的标题信息进行文本分类的一种方法。标题中蕴含了文本的主要信息，是对文本内容的高度概括，并且标题有着简洁、语句简单等特征，使得对文本标题的分析更准确有效。葛文镇等(2016)[3]提出了基于层级类别信息的双向特征选择算法，并实现了标题分类，能有效提高分类效率。杨敏，谷俊(2012)[4]构建了基于混合特征矩阵的支持向量机算法对中文书目进行分类的书目自动分类系统。缪建明等(2008)[5]通过标题信息蕴含的领域信息激活对应的HNC领域，对文本进行了自动分类，测试速度、分类的准确度有了明显的提升。

2)按文本内容分类

按文本内容进行分类，是最常见的一种分类，其关注点在于能区别不同文本内容的关键性词语。按文本内容分类是对根据文档主题进行自动分类，在教育学、法学等领域研究中都十分有用。吴昊(2009)[6]阐述了基于Web信息挖掘的英语阅读自动选篇的分类研究方法，将按文本内容分类应用于教育研究中。在法学研究中，熊小梅等(2007)[7]基于LSA的二次降维法以及改进的互信息特征提取通过文本分类对中文法律案情文本进行自动分流，加快了分类系统的处理速度，减轻工作人员的负担。

3)按情感倾向性分类

按情感倾向性分类是指根据文档中作者对所表达的事物所持有的观点、态度，如正面、负面、积极、消极、中性等。在研究中也被称作情感分析、观点挖掘或是文本意见挖掘等。按情感倾向性的分类中，情感特征的选择与抽取对分类的性能有比较大的影响。目前的文本情感分析在网络舆情分析、政策文件分析、问卷调查等方面应用较多。朱建平等(2016)[8]结合词云、关联规则、文本倾向性分析等技术对中国房地产网络舆情做了实证分析与研究并给出了相关的政策建议。邓雪琳(2015)[9]采用文本分析法，使用文本挖掘软件对政府文件中的关键词等做计量，测量了中国政府职能的转变并给出了中国政府职能的转变逻辑和发展趋向。张珍珍，李君轶等(2014)[10]主要采用问卷调查与网络评论的方式获取游客对旅游的形象感知数据，通过文本挖掘对比分析研究两种途径的旅游形象问题。夏火松等(2013)[11]通过对商品具体属性情感倾向的分析将情感分类分为初分类和细分类两个过程，构建的细分类模型缩短了购买决策的时间，降低了决策的复杂度，并经过特征算法的测试发现情感细分类中互信息达到了更高的准确度。

4)按文本风格分类

按文本风格分类主要是指在文本语言特色方面的分类，是对文本作者在词语使用、句式使用等方面的特色进行分类。针对这种分类方式可以应用于文本作者身份识别、文学作品流派等的研究中。施建军(2011)[12]运用支持向量机技术对《红楼梦》进行了分类研究，更有效地区分了古典文学作品的作者。年洪东，陈小荷等(2010)[13]利用支持向量机统计机器学习模型对中国现当代文学作品做了作者身份识别研究。张云良等(2009)[14]利用向量空间模型及混合句类分解等技术构建了作者写作风格分类器研究了《红楼梦》的作者问题。

3 常见文本分类算法

分类是数据挖掘中的重要方法之一，在文本分类中常见的算法有朴素贝叶斯算法、支持向量机、K近邻算法、Rocchio算法等。朴素贝叶斯算法是在文档自动分类中应用概率模型的一种简单而有效的方法，关注的是文档属于某类别的概率。支持向量机是通过构造一个分类超平面，使得分类间隔达到最大，最大限度地分开两类训练样本的一种方法。K近邻算法是为待分类文本找出最为相似的K个样本，统计这些样本所属的类别，待分类文本的类别就是包含样本最多的类别。Rocchio算法是对一个类别里的所有样本文档各项计算平均值，得到一个称为质心的新向量，若需要对新文档作判断时就通过计算距离比较新文档和质心的相似程度。下面主要对朴素贝叶斯算法、支持向量机、K近邻算法、Rocchio算法等四种算法进行比较分析，见表1所示。

表1 四种分类算法的比较分析

算法	优点	缺点
朴素贝叶斯算法	算法比较简单，在特征属性相关性较小时具有最优的性能，对缺失数据不太敏感，需估计的参数较少	不能对类概率做出非常准确的估计，在属性较多或属性间相关性较大时效率较低
支持向量机	使用的训练集少，可处理高维稀疏文本数据，对特征相关性不敏感	过于依赖分类面周围的正例和反例的位置，核函数的选择缺乏指导，当样本较多时训练速度较慢
K近邻算法	不需要特征选取和训练，易处理类别数目多的情况，方法简单且性能稳定	样本量较大，空间复杂度高，计算开销大于其他方法，K值的选取也直接影响着分类的性能
Rocchio算法	算法容易实现，已被理解，效率高	受文本集分布的影响计算出的中心点可能落在相应的类别外

4 结束语

本文对文本分类及常见的分类算法作了综述性研究，阐述了文本分类的发展和一般流程，根据文本标题、文本内容、文本情感倾向性以及文本风格等角度对文本进行了分类，最后针对几种常见的分类算法作了各自优缺点的比较分析，希望对读者研究文本分类及算法有一定的参考价值。

参考文献：

[1] 王仁武. Python与数据科学[M]. 上海: 华东师范大学出版社, 2015: 267.
[2] 陆旭. 文本挖掘中若干关键问题研究[M]. 合肥: 中国科学技术大学出版社, 2008: 2.
[3] 葛文镇, 刘柏嵩, 王洋洋, 等. 基于层级类别信息的标题自动分类研究[J]. 计算机应用研究, 2016, 07: 2030-2033.
[4] 杨敏, 谷俊. 基于SVM的中文书目自动分类及应用研究[J]. 图书情报工作, 2012, 56(9): 114-119.
[5] 缪建明, 张全, 赵金仿. 基于文章标题信息的汉语自动文本分类[J]. 计算工程, 2008(20): 13-14,17.
[6] 吴昊. 一种Web信息挖掘的英语阅读选篇分类研究[J]. 现代教育技术, 2009, 19(2): 67-70.
[7] 熊小梅, 刘永浪. 基于LSA的二次降维法在中文法律案情文本分类中应用[J]. 电子测量技术, 2007(10): 111-114.
[8] 朱建平, 谢邦昌, 骆翔宇, 等. 中国房地产网络舆情分析[J]. 数理统计与管理, 2016, 35(4): 722-741.
[9] 邓雪琳. 改革开放以来中国政府职能转变的测量——基于国务院政府工作报告(1978-2015)的文本分析[J]. 中国行政管理, 2015(8): 30-36.

(下转第232页)

Subtraction and Frame Difference Based Moving Object Detection for Real-Time Surveillance[J].Journal of Donghua University, 2003, 20(1):15-19.
[11] Mohammed Mahfuz Abdelkadir. Research on Detecting Moving objects using Background Subtraction and Frame Difference[D].黑龙江:哈尔滨工程大学,2012.
[12] 屈晶晶,辛云宏. 连续帧间差分与背景差分相融合的运动目标检测方法[J].光子学报,2014,43(7):1-8.
[13] CHEN Jun-chao, ZHANG Jun-hao, LIU Shi-jia, et al. Improved Target Detection Algorithm Based on Background Modeling and Frame Difference [J].Computer Engineering, 2011, 37:171-173.
[14] 邱联奎,刘启亮,赵予龙,李冠杰.混合高斯背景模型目标检测的一种改进算法[J].计算机仿真,2014,31(5).
[15] Hisilicon.海思Hi3515 H. 264编解码处理器用户指南[M].Revision02.深圳:海思公司,2010
[16] Hisilicon.海思Hi3515媒体处理软件开发参考[M].Revision02.深圳:海思公司,2010
[17] 韦东山.嵌入式Linux应用开发完全手册[M].北京:人民邮电出版社,2008:240-360.
[18] 吴光辉.基于Hi3515的视频传输终端的设计与实现[D].成都:电子科技大学,2012.
[19] 李潺,郭志涛,李伟超,等. 基于Hi3515嵌入式系统的无线车载监控的设计[J].计算机应用与软件,2012,29(9):252-296.

(上接第226页)

[10] 张珍珍, 李君轶. 旅游形象研究中问卷调查和网络文本数据的对比——以西安旅游形象感知研究为例[J]. 旅游科学, 2014(6): 73-81.
[11] 夏火松, 朱慧毅, 魏凤蕊. 商品主观评论的情感细分类模型研究[J]. 情报杂志, 2013(2): 117-120,92.
[12] 施建军. 基于支持向量机技术的《红楼梦》作者研究[J]. 红楼梦学刊, 2011(5): 35-52.
[13] 年洪东, 陈小荷, 王东波. 现当代文学作品的作者身份识别研究[J]. 计算机工程与应用, 2010, 46(4): 226-229.
[14] 张运良, 朱礼军, 乔晓东, 等. 基于句类特征的作者写作风格分类研究[J]. 计算机工程与应用, 2009(22): 129-131+223.