# 第一章 绪论

## 1.1 研究背景与意义

近几年来，随着互联网的蓬勃发展，尤其是移动智能终端的普及，我们的社交、工作、学习、娱乐、出行、购物等都充满了互联网的身影。全世界现在有超过30亿网民，而且这个数量还在不断增加。我国从2013年开始进入4G网络时代，并且规划在10年内进入5G时代，这意味着我们能随时随地使用各种电子设备享受便捷的网络服务。根据2016年6月我国发布的统计报告，我国网民数量已经达到了7.1亿，互联网正在逐渐取代传统媒介称为人们日常生活中获取信息的主要来源。

然而，随着互联网需求的不断增加，网络中的信息量飞速增长。互联网中网页信息80%以上以文本形式存在。文本信息资源作为非常重要的信息类型之一，是人们处理信息资源的重点研究对象之一。想要快速的在互联网中区分和找准信息，就必须对文本信息资源进行分类判断，文本自动分类技术因此得到诞生和发展。文本自动分类是数据挖掘的技术分支之一，作为处理海量信息的重要手段，文本自动分类很大程度上降低查找信息的困难，快速准确地给用户提供想要的信息内容。也就是通过文本自动分类技术，取代传统的、低效的人工分类方法，对文本信息进行快速而准确的分类，提高利用效率。倘若计算机能够帮助人们预处理一些信息的内容判断的工作，那必定会给人们的工作和学习带来非常大的帮助和支持，并将缓解人们当前无法充分利用信息的尴尬情况。

目前，文本自动分类技术已经被其广泛应用，在很多领域都发挥着重要作用，例如搜索引擎、垃圾邮件过滤、主动信息推送服务等。政府可以通过分析公民在社会问题上的在线文本来做出更为公开的决定，公司可以通过分析在线产品评估来识别产品缺陷并预测市场需求，消费者可以通过确定大量在线产品的情绪取向来做出适当的采购决策评论。

因此，针对互联网上海量的、杂乱无章的网页信息资源，利用计算机自动对网页中的文本信息进行分类的必要性就体现了出来，对文本分类技术进行深入研究无论从理论还是实际领域都有重要的意义。

## 1.2 国内外研究现状与发展趋势

文本分类的思想一开始是从信息检索系统演变而来，信息检索是指人们共享不同行业之间的信息时需要提取自己感兴趣的信息。但是，信息的类别十分繁多、冗杂，如何从大量复杂的信息中快速获取有效的信息就变得非常重要。

自动文本分类的研究始于20世纪中期，1958年，美国Luhn提出了采用词频统计来提取摘要的思想，他采用词语的频率与分布信息来估计每个词语的相对重要性。然后再估计每个句子的相对重要度，得分高的句子就被抽取为摘要。1960年，Maron和Kuhns提出了自动关键词分类技术，这是第一篇关于文本分类的文章，对文本分类领域的研究与发展起到了非常深远的影响。他们开创性地提出了朴素贝叶斯分类方法，是一种基于概率的分类方法。在此之后，很多著名的情报学家例如Sparck和Salton等也在文本分类领域取得了卓有成效的研究。

20世纪80年代前，因为技术发展的限制，文本分类大多是基于知识工程的，也就是需要专家人工构建知识工程技术，才能保证分类的有效性。专家制定分类规则是基于知识工程技术分类系统最主要的特点，这些分类规则的重新组合就构成了一个完整的分类系统。由此可以看出，这种分类系统需要大量的专家和知识工程师参与，分类效率和分类准确度都十分低下，而且由于专家研究领域的局限性，这种分类系统的适应性往往也不强。

1990年以来，随着信息技术的快速发展，互联网在人们的生活中占据越来越多的比重，随之而来的信息大爆炸，传统的基于知识工程的分类系统无法处理这种快速增加的数据量。同时，因为机器学习技术的发展，基于机器学习和基于统计的方法正逐渐取代传统方法成为主流的文本分类技术，这也是现在的分类系统的基础。这种分类方法主要通过某种特征选择方法，从训练集中归纳或选择某些特征构成特征集合，以此来创建分类器。这种方法把人们从传统的手工分类中解放了出来，节约了大量的人力物力，成本更低，同时还提高了分类的准确度。

国外的文本分类技术的研究与发展大体可以分为如下三个阶段：第一个阶段进行自动文本分类的可行性与意义的研究，主要时间为1958-1964年；第二个阶段进行实验探索自动文本分类的合理性，主要时间为1965-1974年；第三个阶段进行自动文本分类研究的应用，主要时间为1975-现今；国外的分类系统已经从最初的可行性探索完成到现在的实用化阶段。目前，国外已经有很多可以应于文本分类应用研究的大型商业挖掘软件，例如SAS、SPSS、KXEN等。

我国对于自动文本分类的研究进展要比国外慢很多。一方面，这是因为我国对于文本分类的研究起步较晚。据了解，1981年，南京林业大学的侯汉清教授开启了自动文本分类的探索之路。从基础知识入手，介绍了国外的文本分类技术的发展趋势，清楚地认识到这项技术的重要研究意义。随着技术的发展和研究的深入，中国科学院、清华、北大、哈工大等知名大学都成立了专门研究自动文本分类技术的重点实验室。到目前，我国已经有了基于数据挖掘、大数据分析、机器学习等的文本分类技术。另一方面，因为中文和英文的巨大差异造成的。英文的单词之间有空格分割开来，中文主要是以句分开的，词与词之间则是连续的，没有自然的界限，相比英文分类，中文文本分类要先进行预处理，对连续的词进行分词操作。除此之外，在文本分类中研究英文的语法分析与句法分析的比值要小于中文的语法分析与句法分析之比，这也一定程度上加大了中文文本分类的研究难度。总的来说，我国的文本分类技术的发展也经历了三个阶段：第一个阶段是研究与模仿国外的研究成果；第二个阶段是完善已有的分类技术；第三个阶段是面向中文的发展阶段。

经过几十年的发展，文本分类技术取得了非常大的进步。自动文本分类技术也正在被应用在各个领域，正在影响着人们的生活和学习，有了很好的实践意义和实用的价值，很多优秀的文本分类系统被研究者提出。然而就文本分类而言，仍有可以改进的空间，尤其是特征选择和特征降维方面，还有很多工作可以开展。

## 1.3 本文所做的工作

本文对文本分类技术进行了详细的探讨和研究，包括文本预处理、文本表示模型、特征选择技术、分类器的选择等。由于特征选择技术在文本分类过程中占有着重要的地位，特征选择的好坏直接影响着最终的分类准确度和分类效率，所以本文对特征选择部分进行了详细的介绍，并提出一种混合CHI和MI的改进特征选择方法，克服CHI和MI针对低频词敏感的缺点，提高特征选择的效果。接着本文重点研究了文本分类的算法，介绍主流的文本分类算法，包括朴素贝叶斯算法、K-最近邻算法、决策树算法和支持向量机算法，其中支持向量机算法对海量数据的处理效果最佳，本文也对支持向量机进行重点研究。主要介绍一对一策略下的支持向量机的算法，针对一对一策略算法的只对当前处理的两个类别敏感的缺点，提出使用余弦相似度来计算不同文档之间的相似度，对不同类别文档有所取舍，从而降低类别敏感的影响，最终提高分类的准确性和分类效率。

## 1.4 本文的组织结构

本文主要研究基于机器学习的文本分类技术，所有内容都围绕研究课题展开，具体可分为六个章节，论文的结构以及各章节的主要内容如下：

1. 绪论。主要说明本课题的研究背景与研究意义。接着介绍了本课题的国内

研究现状以及未来的发展趋势，说明国内研究现状及落后原因。最后对本文的组织结构进行了简单的介绍。

第二章 文本分类技术。主要是对文本分类技术的介绍，包括文本分类的基本流程，文本的预处理技术，常用的文本表示模型，以及基本的特征选择技术的介绍，还有常见的分类器选择问题，最后介绍文本分类中方法好坏的评估标准。为后文的研究工作奠定基础。

第三章 混合特征选择方法。主要介绍当前常用的特征选择方法，并分析常见特征选择方法的不足，提出一种改进的混合特征选择方法。其中针对卡方分布特征选择方法和互信息特征选择方法对低频词敏感的缺点，分别引入调节因子，然后混合改进后的两种特征选择方法。

第四章 基于SVM的二分类器处理。介绍SVM中常用的一对一策略，分析一对一策略的缺点，对于不是当前正在处理的两个类别的文档来说，对分类器的影响不是积极的，甚至会影响最终的判断。本文提出了一种使用余弦相似度来判断文档相关程度的方法，去除不相关的文档，提高判断准确率。

第五章 基于机器学习的文本分类系统仿真。通过搭建一个文本分类系统，对前面几章研究的改进方法进行实验验证，分别对比常用特征选择方法和常见的分类器策略与本文研究的方法处理后的分类准确率、召回率和F1值。验证本文研究方法的有效性和实用性。

第六章 总结与展望。针对本文的研究进行了总结和对未来该领域的发展进行了展望。

# 第二章 文本分类技术

## 2.1 文本分类的定义

## 2.2 文本分类的流程

### 2.2.1 文本预处理

### 2.2.2 文本表示及模型

### 2.2.3 特征降维

### 2.2.4 分类模型训练

### 2.2.5 测试方法

### 2.2.6 性能评估

## 2.3 小结