第一章 绪论

1.1. 研究背景

自动文本分类(text categorization，TC)是指在给定的类别体系下对未知类别的文本根据其内容将其自动划归到一个或多个类别的过程。随着信息技术的飞速发展，Internet 上包括维吾尔文在内的多语种信息资源呈现指数级的增长趋势。为了有效地管理和利用这些纷繁杂芜的海量信息，基于内容的文本挖掘技术决定着基于 Web 的信息检索、文本过滤等技术的发展水平。文本自动分类技术在提高信息利用的有效性和准确性上具有重要的现实意义和广阔的应用前景。很多文本处理问题都可以归结为文本分类问题或者需要利用到文本分类技术。例如：信息检索可以看成是文本与查询是否相关的二分类问题，文本过滤(包括垃圾邮件过滤、主题跟踪等)也是二分类问题，自动问答系统需要对问题进行分类，以及 WEB 网页分类、词义消歧、文本自动标签等等。文本自动分类技术是诸多信息处理技术的有机组成部分。20 世纪 90 年代以来得到了长足发展的基于机器学习的文本自动分类方法，在分类模式、算法的性能及其可扩展性、分类效果等方面均有所突破，文本自动分类技术研究已经成为信息检索和文本挖掘领域的研究热点与核心技术[1]。

随着互联网的普及以及近年来，国家在政策、资助等方面给予的大力支持下，我国各民族文化、语言、文字等在民族信息化建设中得到了极大地改善与发展。在这种大好局面下，维吾尔文作为新疆地区主要的少数民族语种之一，在维吾尔文语言文字计算机信息处理，标准化等方面，尤其是维吾尔文 WEB 建设方面得到了迅速的发展，多种类型的维吾尔文网站日益增多，提供内容丰富的 WEB 信息服务。除此之外，越来越多的维吾

尔文文本信息正被以电子文档的形式保存下来，例如，维吾尔文报刊杂志、维吾尔文数字图书馆等等，促进了少数民族地区教育和经济发展，提升了社会的信息化水平。然而，如何有效地管理和利用这些大量的维吾尔文电子文本信息也就成为了一项重要的研究课题。文本自动分类作为文本挖掘领域中的核心技术，能够解决大量文本信息的归类问题，是管理和利用海量文本信息的有效技术手段，有着非常广泛的应用前景。但是迄今为止还没有研究人员对维吾尔文文本分类所涉及到的文本特征表示、特征空间降维、分类器训练等技术进行过比较全面、系统地研究，严重制约了维吾尔文文本倾向性分析、意见挖掘、网络舆情挖掘以及有害文本信息过滤等应用系统的研究开发。

经过国内外研究人员多年的研究，基于英文和中文等大语种的文本分类技术已经相当成熟并且已经进入了实用阶段。然而，对于维吾尔语而言，文本分类相关技术方面的研究仍然还是一个空白。由于文本分类技术自身的特点以及维吾尔文语言文字的特殊性，我们无法直接套用在中英文中常用的技术，特别是维吾尔文文本分类数据集的获取以及预处理、文本特征表示、特征空间降维等相关的关键技术和方法都与具体语言文字特点密切相关。在过去的二十年中，维吾尔语信息处理技术在维吾尔语字母编码设计、字库制作、输入法研发、信息系统本地化以及词干提取、词性标注等方面取得了一定的突破，为进一步研究基于维吾尔文的文本信息处理相关技术奠定了基础[2]。

深入研究维吾尔文文本自动分类关键理论及技术方法，结合维吾尔文语言文字特点，建立维吾尔文文本数据集以及预处理方法，构造适合维吾尔文的文本特征表示以及特征空间降维方法，进而建立起高效的维吾尔文文本分类模型为基于维吾尔文的信息检索、文本过滤，维吾尔文文本倾向性分析、网络舆情分析以及话题识别等应用系统的研究开发奠定基础，有着非常重要的理论意义和实用价值。

1.2. 研究意义

文本分类作为一项具有很高实用价值的关键技术，是组织和管理文本信息的有力手段。近年来，文本分类在信息过滤、信息组织和管理、词义辨析和数字图书馆等领域得到了广泛的应用，并且取得了很大的发展[3]。下面，就对文本分类在上述几个领域的应用进行介绍。

1.2.1. 信息过滤

随着互联网的迅猛发展和普及，人们可以在网络上便捷地获取越来越多的信息。然而，所获取的信息量过大会给人们对信息的处理带来很大困难，不仅无法快速找到自己所需要的信息，有时还会带来一些负面的信息。于是，人们迫切需要根据自己的需求，对源源不断获取的信息进行动态过滤，保留相关信息，屏蔽无关信息，这就是所谓的“信息过滤”。从文本分类的角度来讲，它将所有文本分为“相关文本”和“无关文本”，属于两类文本分类问题。

信息过滤能够主动地获取用户特定的信息需求，使用这些信息需求组成过滤条件，进而对信息资源进行过滤，这样就可以把符合条件的信息过滤出来进行服务。因此，信息过滤具有两个显著特点:个性化和主动化。个性化的实质是针对性，即针对不同的用户采用不同的服务策略，提供不同的服务内容；主动化的实质是主动性，即无需用户做什么事情，系统自动按照用户的需求提供相应的服务。信息过滤的个性化和主动化将使用户付出尽可能少的努力，却可得到尽可能好的服务。

在传统的信息获取技术中，用户是主动方，因此可以称之为“拉”(Pull)，与之相反的另一种方式则称为“推”(Push)，由信息发布方向感兴趣的用户主动推送信息。用户的喜好可以由用户自己设置，或者利用用户访问过的信息集合进行描述。面对用户形式各异的个性化信息需求，完善的信息过