基于深度学习的多标签文本分类的研究

一. 成果完成人及所在单位：吴杰隆，光电与通信工程学院，指导教师：张宏怡

二. 成果简介：

本课题设计基于 Python 语言的爬虫数据收集模块，从知乎上获取大量数据，并采用 Word2Vec 工具训练词向量以便于文本表示。同时设计基于深度学习的多标签文本分类模型获取数据处理的高效性。设计一个高性能,高精度的多标签自动标注系统对于降低知乎网站的运营成本就显得尤为重要。实现对问题和内容的自动化打标签,可以有效地解决标签与问题不相关的问题和节约大量人力成本,并且知乎网站可以更精确地向用户推荐内容、提高内容的分发效率、满足用户对信息资源的需求和提升用户体验。

三. 主要研究内容及创新点：

该系统主要分为三个模块，分别为数据收集模块，数据预处理模块，文本分类模块。其中，数据收集模块负责数据的收集,该模块主要由爬虫程序构成,爬虫程序负责从知乎网站爬取文本数据,并设计一些规则来筛选爬取的数据,最后将文本数据存到计算机本地文件中,方便数据预处理模块的调用。数据预处理模块。该模块的输入数据是数据收集模块收集的数据。数据预处理模块主要负责清洗爬取的文本数据,筛选标签,对数据进行统计分析,将文本转为计算机可以计算的形式,最后将数据输入到文本分类模块。文本分类模块。文本分类模块主要负责对输入的文本数据进行分类,实现自动化打标签。设计和实现数据收集模块和数据预处理模块。设计并实现了基于 Python 语言的爬虫数据收集模块，设计并实现数据预处理模块。设计和实现基于深度学习的多标签文本分类模型,具体包括:基于 CNN、LSTM、CNN-LSTM 的分类模型。

首先，解决了关于文本预处理中文本分词采用何种算法的问题。分词算法主要分为两类：基于字符串匹配的算法和基于统计的机器学习算法。基于字符串匹配的方法简单高效，但对未登录词的分词效果较差，词典的大小和质量也会影响分词效果和性能。基于统计的机器学习分词算法能很好的处理歧义词和未登录词,效果好于基于字串匹配的算法。所以，采用机器学习分词算法解决了关于文本预处理中文本分词采用何种算法的问题。

其次，解决了关于文本表示中采取何种表示方法的问题。文本表示一般会采用以下 2 种形式：布尔编码表不和分布式表征。其中布尔编码表示即简单又直观，但是也存在一个严重的问题，就是任意两个词之间都是孤立的，无法表征语义信息，而且当词库量较大时，向量会具有高纬度和高稀疏性,而这会造成计算量大和占用大量内存。分布式表征是一种既可以表示词本身又包含语义信息的低纬度的表示方法。我们采用分布式表征（词向量）中的词向量训练算法 Word2Vec 中的 Skip-gram 模型。

最后，解决了关于如何构建分类模型的问题。目前,典型的神经网络结构有深度神经网络、卷积神经网络、循环神经网络等，这些神经网络结构被广泛应用到自然语言处理、医学、自动驾驶等领域，并获得了极好的效果。所以本文采用卷积神经网络、循环卷积神经网络中的长短期记忆网络构建分类模型。

四. 主要研究成果及应用情况

申请（授权）专利 1 项。

联系人：吴杰隆，　　电　话：15610318596，　　Email:15610318596@163.com