甘肃省大学生创新创业训练计划

项目申报表

（创新训练项目）

推荐学校：	西北师范大学
项目名称：	基于改进特征空间的短文本表示方法
所属一级学科名称:	计算机科学与技术
项目负责人：	孔佳睿
联系电话：	18419616964
指导教师：	马慧芳
联系电话：	18009481665
申报日期：	2018 年 4 月

甘肃省教育厅 制

二〇一八年四月

<table>
<tr><td colspan="2">项目名称</td><td colspan="7">基于改进特征空间的短文本表示方法</td></tr>
<tr><td colspan="2">项目所属
一级学科</td><td colspan="7">计算机科学与技术</td></tr>
<tr><td colspan="2">项目实施时间</td><td colspan="7">起始时间：2017 年 12 月　　完成时间：2018 年 10 月</td></tr>
<tr><td>项目简介
(100字以内)</td><td colspan="8">本课题主要贡献有：从类内及类间两个角度，提出改进的特征权重分数进行特征选择；选择各类中特征权重分数靠前的特征词优化特征空间，构建词语相似度矩阵；利用词语相似度矩阵，将不同长度的短文本映射成同维向量。</td></tr>
<tr><td rowspan="5">申请人或申请团队</td><td></td><td>姓名</td><td>年级</td><td>学号</td><td>所在院系/专业</td><td>联系电话</td><td colspan="2">E-mail</td></tr>
<tr><td>主持人</td><td>孔佳睿</td><td>2015级</td><td>2015710 30213</td><td>计算机科学与工程学院/计算机科学与技术（非）</td><td>1841961696 4</td><td colspan="2">1786756024@qq.com</td></tr>
<tr><td rowspan="3">成员</td><td>王军艳</td><td>2015级</td><td>2015710 302</td><td>计算机科学与工程学院/计算机科学与技术（非）</td><td>1739315085 9</td><td colspan="2">1787345308@qq.com</td></tr>
<tr><td>马春燕</td><td>2016级</td><td>2016710 20116</td><td>计算机科学与工程学院/软件工程</td><td>1519315786 7</td><td colspan="2">2174806498@qq.com</td></tr>
<tr><td>王文涛</td><td>2016级</td><td>2016710 20127</td><td>计算机科学与工程学院/软件工程</td><td>1869311172 5</td><td colspan="2">1092988384@qq.com</td></tr>
<tr><td rowspan="3">指导教师</td><td rowspan="2">第一指导教师</td><td>姓名</td><td colspan="2">马慧芳</td><td>单位</td><td colspan="2">西北师范大学</td></tr>
<tr><td>年龄</td><td colspan="2">37</td><td>专业技术职务</td><td colspan="2">副教授、副系主任</td></tr>
<tr><td colspan="2">主要成果</td><td colspan="6">主持在研国家自然科学基金 2 项、参与国家自然科学基金 3 项。在《Information sciences》、《Neurocomputing》、《Knowledge and Information Systems》、《软件学报》、《电子学报》、IJCNN、PAKDD、PRICAI 等国内外重要期刊和国际知名会议上发表论文 40 余篇，其中 SCI 收录 4 篇，EI 收录 18 篇。担任《Artificial Intelligence Review》、《IEEE Transactions on Cybernetics》、《IEEE Transactions on Systems、Man and Cybernetics》等期刊的审稿人。</td></tr>
</table>

一、申请理由（包括自身具备的知识条件、自己的特长、兴趣、已有的实践创新成果等）

（1）团队基本情况介绍

本项目组的人员包括不同年级结构层次合理，每个人都有相应的优点，在探讨中能够从不同的角度思考，提供不同的解决问题的思路。更重要的是，项目组的每个成员都同在“社交网络环境下短文本信息处理研究”科研创新团队里面，接受了将近一年的短文本挖掘的相关培训，也做了一定的前期的工作。

（2）数据获取及实验可行分析

由于很多相关学者在文本挖掘方面做研究，所以可以获取大量的可供参考的资源。同时网上大量的公测数据集是可以获取下来进行实验的，所以这又为实验验证算法有效性提供了可能。短文本的预处理，在网上有很多成熟的代码，而我们所要关心的就是算法的实现，方法的优化改进。

二、项目方案

具体内容包括：

1、项目研究背景（国内外的研究现状及研究意义、项目已有的基础，与本项目有关的研究积累和已取得的成绩，已具备的条件，尚缺少的条件及方法等）

国内外的研究现状：

目前大多的文本表示模型都建立在 ***VSM*** 的模型之上的，但是该模型忽略了文本词语间的上下文语义信息，事实证明简单考虑词语的共现关系在很多长文本的处理中已经能够取得不错的效果。由于短文本自身特点，往往致使数据集矩阵出现高维稀疏的问题。特征集的降维操作已成为文本分类领域重要的研究课题之一。其中，降维的核心问题是特征选择。

一般地，特征选择过程如图 1 所示。特征选择从原始特征集中选出最具分类能力的特征生成特征子集，利用相关评价准则不断筛选，直至特征子集满足条件，最后验证选择结果的好坏，特征子集中的词保留了原始数据集中的大部分信息，同时具有很强的分类贡献能力。

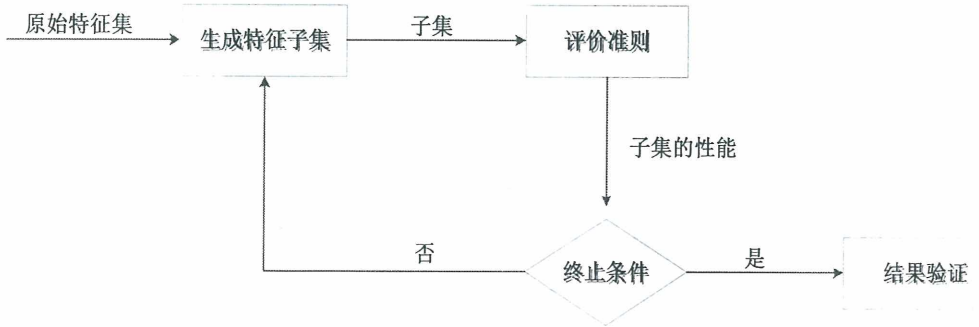

图 1　特征选择算法结构图

特征选择过程中最关键的是评价函数的设计。常用的特征子集评价方法有文本频率(Document Frequency,DF)、信息增益(Information Gain,IG)、卡方检验(CHI-square test,CHI)、互信息(Mutual Information,MI)、期望交叉熵(Expected Cross Entropy,ECE)等。特征子集的好坏判定也是特征选择算法的研究热点之一。Monica 等人通过综合考虑 DF、IG、CHI 等特征选择方法，指出组合后的特征选择方法分类效果更佳；Zhang 等人将基于正则化的互信息和类的分布信息应用于文本分类；Liu 等人将正则化引入传统互信息和信息增益方法进行文本的特征选择，取得了不错的效果；Karl 提出一种基于 KL 散度（Kullback-Leibler divergence，KLD）的特征选择评价函数；徐燕等人提出了基于区分类别能力的高性能特征选择方法，并依此构造了知识增益特征选择函数,实验验证了算法的有效性；张延祥等人提出一种在不平衡数据集上进行文本分类特征选择方法，Liu 等人提出基于词性的特征选择算法同时用知网进行特征扩展，实验表明具有非常好的分类效果。

研究意义：

（1）理论意义

本课题在原始的无法从特定类别进行特征选择，以及文本表示直接作用于短文本会出现数据集高维稀疏等问题的传统方法，类内分散度、类间集中度的基础上，研究最适合短文本表示的方法，挖掘短文本的语义及类别信息，充分表示短文本自身蕴含的信息。

（2）现实意义

将不同长度的短文本映射成包含一定类别信息的同维向量，优化短文本的表示方法，对短文本的聚类、分类以及相似度计算等的研究都有重要的现实意义。

2、项目研究目标及主要内容

本课题项目研究的目标是：在信息增益的基础上，通过从类内、类间角度充分考虑特征词的类别信息，定义特征权重分数计算方法，提出一种改进的特征选择方法来降低数据集维度、抽取具有较强类别信息的特征词，构建优化的特征空间优化短文本的表示形式，不再仅仅只是简单的用词语来表示文档。

研究内容：

◆ 短文本特征选取的研究

短文本的特征选择方法的优化是本课题的重点，传统长文本的处理中，经典的向量空间模型完全忽略了文本词语间的上下文语义信息，事实证明简单考虑词语的共现关系在很多长文本的处理中已经能够取得不错的效果。但是对于短文本的特征选取来说，必须用到文档频率法(*DF*)、互信息(*MI*)、信息增益(*IG*)、期望交叉熵(*ECE*)等方法。优化了短文本的表示形式，不再仅仅只是简单的用词语来表示文档。

◆ 基于改进特征空间的短文本表示研究

文本分类中，高维稀疏的特征空间往往直接影响分类的性能，为提高分类的准确率和效率，特征集的降维操作已成为文本分类领域重要的研究课题之一。其中，降维的核心问题是特征选择。文本特征选择通常采用文本特征评价函数计算词语相关于类

别的权重，按词语权重值进行排序和筛选得到特征词子集，因此特征词子集的质量由特征评价函数直接决定。本课题将在信息增益的基础上，从类内和类间两个角度考察，充分考虑特征词出现与不出现的情况和特征词的类别分布情况，在从训练词集中选择具有强类别代表能力、强类别区分能力的高质量特征词，优化特征空间及短文本的表示。

拟解决的关键问题：

- ◆ 如何寻找合适的方法，用于短文本的特征选取。
- ◆ 如何设计算法得到所需的特征权重分数。
- ◆ 如何由已得到的特征权重分数来选取短文本的特征词。
- ◆ 如何通过特征词得到词语相似度矩阵。

3、项目创新特色概述

◆ **特征空间的优化。**

利用类内分散度及类间集中度来定义特征权重分数，从各类别中选择一定数量分数靠前的特征词，作为候选特征子集；然后分别从各类别中选定一个最具代表能力的特征词生成最终的特征空间，构建词语相似度矩阵。

◆ **短文本表示的优化。**

通过词语相似度矩阵，优化短文本表示模型，映射处理后的短文本均为同维向量，强化了短文本包含的类别信息，同时改善了 TF-IDF 权重方法数据集稀疏的问题，优化了短文本的表示。

4、项目研究技术路线

在短文本的特征空间优化中，某一特征词在某一特定类别中的各个文本中分布越均匀，该特征词对该类别的类别代表能力越强，**称该特征词在该类别中具有较强的类内分散度。**某一特征词在某一特定类别中均匀出现，而在其他类别中出现的频数越小，该特征词的类别区分能力越强，**称该特征词具有较强的类间集中度。**在类内分散度和类间集中度的基础上，定义特征权重分数。特征权重分数衡量的是特征词在特定类别中的重要程度，特征权重分数的值越大，表明特征词能提供的类别区分信息就越丰富，越有利于短文本的表示和特征空间的构建。通过特征权重分数得到优化后的特征空间，将不同长度的短文本映射成包含一定类别信息的同维向量，可以有效解决短文本向量高维、稀疏的问题。

（2）技术路线

图 2 给出了改进特征空间方法的总流程图，首先，从类内及类间两个角度充分挖掘特征词包含的类别信息，提出了改进的特征权重分数进行特征选择，选择原始特征集中具备强类别信息的特征词生成特征子集；其次，选择各类中特征权重分数靠前的

特征词优化特征空间，构建词语相似度矩阵，一定程度上解决了原始特征空间高维稀疏的问题；最后，利用词语相似度矩阵，优化短文本表示模型，将不同长度的短文本映射成包含强类别信息的同维向量，且包含足够的类别信息，构造出短文本表示模型。

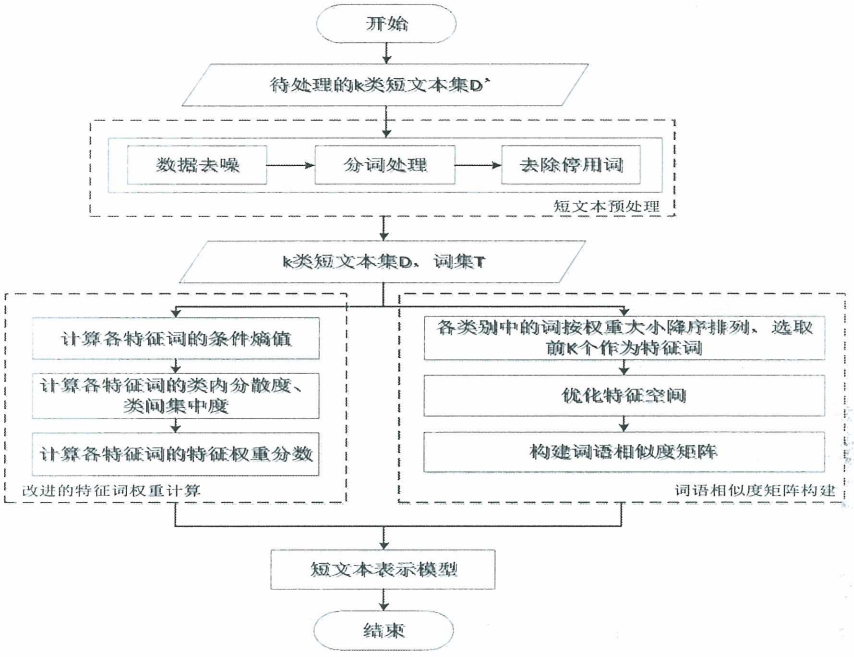

图 2：算法流程图

利用上述算法，最终可以将从背景语料库中挖掘出来的词语语义信息构造成矩阵的形式，然后扩展到短文本的特征空间。解决了短文本在特征上高维稀疏的问题。

5、研究进度安排

（文献查阅）：　2017 年 12 月至 2018 年 4 月

（社会调查）：　2018 年 1 月至 2018 年 2 月

（方案设计）：　2018 年 2 月至 2018 年 3 月

（实验研究）：　2018 年 3 月至 2018 年 5 月

（数据处理）：　2018 年 3 月至 2018 年 4 月

（研制开发）：　2018 年 5 月至 2018 年 6 月

（撰写论文或研究报告）：　2018 年 6 月至 2018 年 8 月

（结题和答辩）：　2018 年 9 月至 2018 年 10 月

（项目鉴定）：　2018 年 10 月至 2018 年 11 月

（成果推广或论文发表）：　2018 年 10 月至 2018 年 11 月

6、项目组成员分工

（1）数据收集：孔佳睿，王军艳；

（2）实验验证：孔佳睿，王军艳；

（4）系统开发：马春燕，王文涛；

（5）相关文档撰写：孔佳睿，王军艳。

三、学校提供条件（包括项目开展所需的实验实训情况、配套经费、相关扶持政策等）

西北师范大学计算机科学与工程学院始终重视实验室建设工作，为教师、研究生和优秀本科生提供良好的科研环境，创造良好的研究氛围。现有实验室总使用面积4000多平方米，设备数量达3000多台套，其中服务器和微型计算机2000余台，设备总值1800多万元。

本次项目的申报来自于学校创新创业学院以及计算机学院领导大力推行的“本科生科研能力提升计划”。

四、预期成果

1.掌握“短文本表示”的整套流程，及其领域中常用的技术；

2.录用一篇国家核心期刊论文；

3.根据本项目提出的“短文本表示”方法做出研究性软件，并申请软件著作权。

五、经费预算

总经费（元）	6000	财政拨款（元）	5000	学校拨款（元）	1000

注：总经费、财政拨款、学校拨款由学校按照有关规定核定数目进行填写

具体包括：

1、调研、差旅费：500元；

2、用于项目研发的元器件、软硬件测试、小型硬件购置费等：2000元；

3、资料购置、打印、复印、印刷等费用：500元；

4、学生撰写与项目有关的论文版面费、申请专利费等：3000。

六、导师推荐意见

本项目针对现有短文本表示方法的不足，提出了一种改进特征空间的短文本表示方法，该方法可以降低数据集维度、抽取具有较强类别信息的特征词，有特色，有创新。团队成员包括不同年级结构层次合理，每个人都有相应的优点，且成员已接受了将近一年的短文本挖掘的相关培训，也做了一定的前期的工作，同意推荐，建议资助。

签名：马慧芳

2018 年 4 月 26 日

七、院系推荐意见

同意推荐

院系负责人签名：

学院盖章：

2018 年 4 月 27 日

八、学校推荐意见：

同意推荐

学校负责人签名：

学校公章

2018 年 4 月 28 日

注：表格栏高不够可增加。