

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

融合语义与图结构的关键词

项目名称：

提取算法应用

所属一级学科名称：

计算机科学与技术

项目负责人：

马兰

联系电话：

15002504374

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

项目名称		融合语义与图结构的短文本关键词提取算法					
项目所属一级学科		计算机科学与技术					
项目实施时间		起始时间： 2017 年 11 月 完成时间： 2018 年 10 月					
项目简介 (100字以内)	针对现有的短文本关键词提取算法未充分考虑词语间的隐含语义及图的结构特征本项目提出了一种融合语义与图结构的短文本关键词提取算法，从海量的短文本数据中快速并准确地获取有用的信息。通过设计实验来验证方法的有效性。						
申请人或申请团队		姓名	年级	学号	所在院系/专业	联系电话	E-mail
	主持人	马兰	2015 级	201571030316	计算机科学与技术	15002504374	3216089290@qq.com
	成员	刘晓倩	2014 级	201471010317	物联网工程	18809442719	1475973011@qq.com
		沈玉红	2014 级	201471010328	物联网工程	18893717915	1915687247@qq.com
		糟智明	2015 级	201571010141	物联网工程	18393711581	1101009881@qq.com
		伍诗萌	2016 级	201671020126	软件工程	18973491639	690141474@qq.com
指导教师	指导教师	姓名	马慧芳		单位	西北师范大学计算机科学与工程学院	
		年龄	37		专业技术职务	副教授	
	主要成果		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。				

一、申请理由

本人在“社交网络环境下的短文本信息处理”科研创新团队里，已接受了两年的相关培训，前期参与的“创新能力提升计划”项目在计算机科学与工程学院第二届挑战杯比赛中荣获一等奖。

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。项目组的每个成员都同在“社交网络环境下的短文本信息处理”科研创新团队里面，接受了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。在前期的工作中，项目组已经对短文本的相关技术有了一定的了解。相信前期的基础工作会给本课题的研究带来一些帮助

二、项目方案

1、项目研究背景

(1) 项目研究意义

当今社会离不开信息交流，随着信息科学技术的快速发展，社交媒体已被公认为是继报刊、广播、电视之后的第四类媒体。社交媒体是人们彼此之间用来分享意见、见解、经验和观点的工具和平台，现阶段主要包括社交网站、博客、微博、视频分享、书签等等。社交媒体在互联网沃土上的蓬勃发展，使得网络文本开始呈现出爆炸式增长，这些文本一般长度比较短，我们称之为短文本(Short Text)，因此网络数据中很大一部分是长度很短的文本数据。各种形式的短文本已经成为我们都能够接收的信息沟通渠道和情感交流手段。短文本数据量异常庞大，数据中包含人们对社会各种现象的各种立场和观点，话题涉及政治、经济、文化、军事、娱乐、生活等各个领域，面对如此庞大的文本数据，人们通常需要花费大量的时间和精力对其进行挑选和甄别，因此如何高效准确地实现关键词提取成为影响信息检索系统性能的关键。本项目可以实现有效地提取短文本的关键词。

(2) 国内外研究现状分析

关键词指能够反映原文的主旨信息的词语，是对文章主题信息的精炼。国内外研究者们提出的关键词提取算法，主要可以概括为三类：基于统计信息的关键词提取算法、基于机器学习的关键词提取算法和基于语义的关键词自动算法。不同的关键词提取方法具有各自不同的优缺点。

▪ 基于统计的关键词提取算法

主要从词频、位置等角度考虑关键词的重要性，常见的算法有 TFIDF 算法^[1]、N-Gram 算法^[2]。该算法比较简单快速，对出现频率较高或者位置较特殊（如段首、段尾、首行等）的词汇能够快速提取，但极易忽略那些出现频率不高且不在重要位置出现但对文章的主题思想和内容具有重要意义的词语，结果具有片面性，准确率不高。

▪ 基于机器学习的算法

通过对训练集进行训练，得到相关模型的特征，然后利用该模型对关键词进行提取，即将关键词提取过程转换成分类过程^[3]，常见的有基于 SVM^[4]的方法、基于朴素贝叶斯^[5]的方法等。该类方法提高了关键词提取的准确率，易于推广，但是依赖于选取的算法模型，训练耗时长。

▪ 基于语义的算法

采用语义词典（如英文的 WordNet^[6]，中文的《同义词词林》、《知网》等）或者词汇链方法来获取词汇间的语义知识来提取文本关键词。该算法提高了提取的准确率，但是依赖于背景知识库、词典、词表等，无法提取出不包含于知识库的词或词组，对文本格式要求严格

（3）可行性分析

为了验证算法的有效性，本项目选取中英文论文标题作为实验数据。由于前期项目组成员已在短文本特征提取方面开展了相应研究，收集了一定规模的论文标题数据，使得实验验证具有了可行性。在实验设计方面，首先设计内部耦合关系和外部耦合关系权重的参数的最佳值，使算法达到最优的状态；然后实验比较本文方法与其他关键词提取方法的性能。目前考虑对语义耦合不考虑图结构特征的关键词提取方法、只考虑图结构特征但不考虑语义耦合的关键词提取方法与本文方法得到的关键词集合进行比较，验证使用本文所提出的关键词提取算法的有效性

2、项目研究目标及主要内容

（1）研究目标：

由于短文本内容较少、错误拼写、不规范用语和噪音比较多、实时性特别强等独特特征，传统的长文本关键词提取方法不再适用于短文本，那么如何从字数如此少的文本中抽取关键词并使其能够准确描述话题主旨是目前关键词提取任务面临的新的挑战。本项目主要研究短文本关键词的提取问题，提出一种可行有效的短文本关键词提取算法。

（2）研究内容：

▪ 文本图构建

文本图中顶点集对应词项集，边集提供了词项间关系应遵守的某些标准，如共现关系。

▪ 顶点权值初始化

构建文本图之后，计算顶点重要性的权值。

▪ 文本图中边的权值计算

一方面，利用词语间内外部语义耦合关系计算词语间的相似度对文本图中的边加权，另一方面，根据所建文本图的结构特征对边加权。

▪ 关键词提取

将两种边的加权方案有效地综合起来进行随机游走迭代计算出节点的重要性，并降序排序取出前 K 项作为初步的文本集关键词集合，然后，根据关键词特征进一步优化，得到文本集最终的关键词排序结果。

3、项目创新特色概述

（1）本项目利用图进行文本表示，文本图中顶点集对应词项集，边集提供了词项间应遵守的某些标准，充分地考虑到了文本的语义信息，使得提取出的关键词更准确有效；

（2）本项目在计算词项间的相似度时，有效地将词语的语义耦合关系和文本图的结构信息结合起来，既考虑到文本的语义信息，又考虑到文本图的内部结构信息，使得在随机游走的过程中，节点之间不再是等概率跳转，而是增大相似节点之间跳转的可能性；

(3) 本项目采用词频和词长两种关键词特征对随机游走迭代结束后提取出来的关键词集合进行进一步优化，得到最终的关键词集合，增强了提取结果的准确性。

4、项目研究技术路线

如图 1 所示，本项目由如下几个步骤组成：

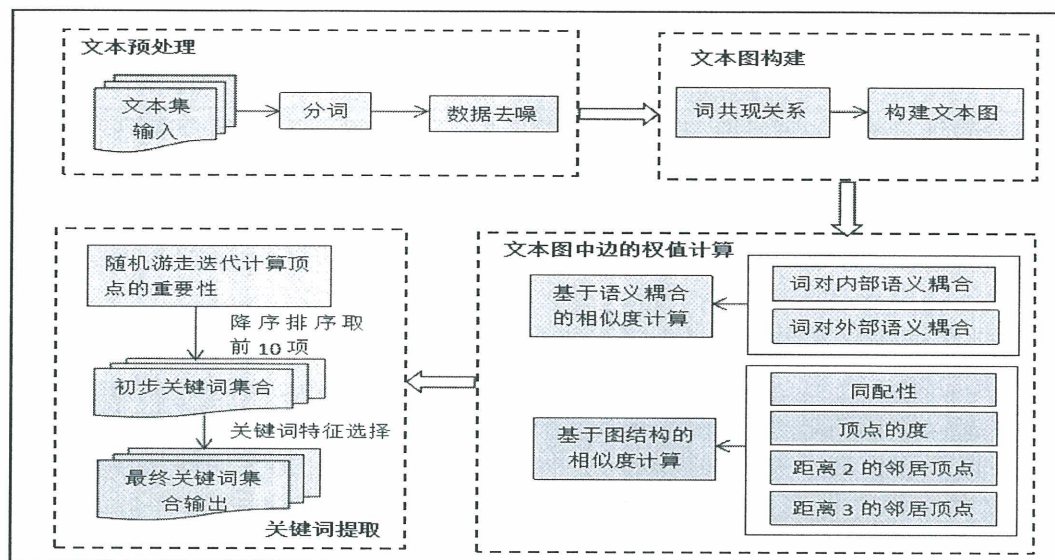


图 1 基于语义耦合和结构相似度的关键词提取算法

- 进行文本预处理，包括分词、词性标注等；
- 根据所有邻边规则构建文本图、根据词性对文本图的顶点权重进行初始化；
- 利用词语间内部耦合及外部耦合关系以及文本图的结构特征计算词语间的相似度作为文本图中边的权值；
- 将两种边的加权方法有效地综合起来进行随机游走迭代提取出初步的文本集关键词集合；
- 根据关键词特征进一步优化，得到文本集最终的关键词排序结果。

5、研究进度安排

（文献查阅）：	2017 年 12 月至 2018 年 04 月
（社会调查）：	2017 年 12 月至 2018 年 01 月
（方案设计）：	2018 年 01 月至 2018 年 03 月
（实验研究）：	2018 年 03 月至 2018 年 05 月
（数据处理）：	2018 年 01 月至 2018 年 03 月
（研制开发）：	2018 年 03 月至 2018 年 07 月
（撰写论文或研究报告）：	2018 年 07 月至 2018 年 08 月
（结题和答辩）：	2018 年 09 月至 2018 年 10 月
（项目鉴定）：	2018 年 09 月至 2018 年 10 月
（成果推广或论文发表）：	2018 年 09 月至 2018 年 10 月

6、项目组成员分工

刘晓倩：研究相关理论
沈玉红：相关数据处理
糟智明：搜集相关数据
伍诗萌：搜集相关数据

三、学校提供条件

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。依托这些实验室，教师和学生可以开展各类科研项目，创造科研成果，使学生有机会和条件在具体的科研项目中提高能力，培养科研素质。学院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”，共获得学校拨款 20 万元；同时，项目的开展依托于优秀的老师细心指导，及物联网实验室，来进行传感器等物联网工程关键技术研究、相关应用系统开发，为项目的开展提供了有力的设备基础及技术保障。

四、预期成果

1. 掌握“关键词提取”的相关算法；
2. 录用 CSCD 正刊论文 1 篇；
3. 根据本项目提出的关键词提取算法做出相关软件，力争申请软件著作权；
4. 总结研究成果，提供一套相关的理论分析与技术应用方案。

五、经费预算

总经费（元）	6800	财政拨款（元）	5000	学校拨款（元）	1800
具体包括：					
1、调研、差旅费：1000 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：2200 元；					
3、资料购置、打印、复印、印刷等费用：600 元；					
4、学生撰写与项目有关的论文版面费、申请专利费等：3000 元。					

六、导师推荐意见

本项目面向现有短文本关键词提取领域，拟设计融合语义与图结构的短文本关键词提取算法，算法思路清晰可行。此外，团队成员包括不同年级学生，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



院系负责人签名：



学院盖章：



2018 年 4 月 27 日

八、学校推荐意见：



学校负责人签名：



学校公章



2018 年 4 月 28 日