 DOI: https://doi.org/10.12345/jxffcxysj.v7i3.16265

Analysis and Application of Undergraduate Thesis Titles and Comments Based on Text Mining—Taking Computer Science and Technology as an Example

Xiaohuan Yang Zihui Jin Shixia Zhang

Tianfu College of SWUFE, Mianyang, Sichuan, 621000, China

Abstract

Graduation thesis is an important part to test the comprehensive quality and innovation ability of students, and also an important manifestation of teaching quality in colleges and universities. This paper takes the undergraduate graduation thesis of computer science and technology major from 2019 to 2023 in Tianfu College of SWUFE as the research object, and uses text mining technology to conduct in-depth analysis of the thesis topics and comments. Through data preprocessing, feature extraction, model training and other steps, clustering and sentiment analysis of thesis topics and comments are realized. The results show that text mining technology can help colleges and universities understand students' knowledge ability, research direction and research level, and also can reveal the common problems existing in the thesis, providing a useful reference for teaching reform in colleges and universities.

Keywords

graduation thesis; text mining; teaching reform

基于文本挖掘的本科毕业论文题目和评语的分析与应用——以计算机科学与技术为例

杨晓欢 靳紫辉 张仕霞

西南财经大学天府学院，中国·四川 绵阳 621000

摘　要

毕业论文是检验学生综合素质和创新能力的重要环节，也是高校教学质量的重要体现。论文以西南财经大学天府学院2019—2023届计算机科学与技术专业的本科毕业论文为研究对象，采用文本挖掘技术对论文题目和评语进行深入分析。通过数据预处理、特征提取、模型训练等步骤，实现了对论文题目和评语的聚类和情感分析。结果表明，文本挖掘技术可以助益高校了解学生的知识能力、研究方向以及研究水平，也可以从中透视论文存在的普遍性问题，为高校教学改革提供有益的参考。

关键词

毕业论文；文本挖掘；教学改革

1 引言

本科毕业论文是检验学生综合素质和创新能力的重要环节，也是高校教学质量的重要体现。选题是毕业论文写作的首要环节，既要符合人才培养方案要求，并且能够体现专业特点。毕业论文成绩采用评审和答辩的方式综合评定。指导老师和评审专家分别对学生论文评阅并写评语。通过对毕业论文题目和评阅意见进行分析，能够了解学生的知识能力、研究方向以及研究水平等，这也有助于我们洞察到学位论文存在的普遍问题，为未来论文的撰写和指导提供有益的借鉴[1]。

文本挖掘技术是一种从大量文本数据中提取有用信息的过程，主要包括数据预处理、特征提取、模型训练等步骤。文本挖掘作为一种有效的信息提取技术，已经在多个领域得到广泛应用。在教育领域中，文本挖掘技术可以对教学管理、教学质量监控、课程设计等方面提供有益的参考。例如，有学者采用文本挖掘技术对教学反馈、课程评价、学生评教等方面进行了深入研究。因此，论文基于文本挖掘技术，对论文题目和评语进行主题挖掘和情感分析，以此了解学生研究热点和知识储备，从而为提高本科毕业论文的质量和水平提

【基金项目】四川省民办教育协会（研究中心）2023年科研项目：基于文本挖掘的计算机专业本科毕业论文题目和评语的分析与应用（项目编号：MBXH23YB356）。

【作者简介】杨晓欢（1991–），女，中国云南人，硕士，讲师，从事计算机专业的相关教学与科研研究。

供借鉴。

2 研究对象和研究方法

2.1 研究对象

论文以西南财经大学天府学院计算机科学与技术专业的本科学位论文为研究对象，从论文管理系统中收集 2019—2023 届共计 1042 篇学生的论文题目、学术评语、成绩为主要分析内容。计算机科学与技术专业修订的人才培养方案中，其培养目标侧重培养学生掌握软件项目的分析、设计、开发、测试、运维和管理等相关知识。对学位论文质量的评价采取指导老师和评阅专家共同评阅的措施 [2]。如表 1 所示，是一份 2022 年计算机科学与技术专业的论文评审意见表格。

表 1 论文评审意见

论文题目	指导成绩	指导老师评语	评阅成绩	评阅专家评语
基于 SSM 框架的流浪动物救助平台设计与实现	76	该生对课题系统进行了整体架构及功能模块设计，并完成了整个系统的开发调试，但创新性稍显不足，望在后续研究学习过程中加以改进和提升	70	选题具有较高的实践指导意义，论文内容完整，系统的功能比较齐全，但存在以下几个问题：①摘要翻译不够规范；②数据库设计中的 ER 图实体不完整，缺少用户实体

2.2 研究方法

论文通过 Python 编程，提取 1042 篇毕业论文的题目和评语，基于中文文本挖掘技术，提取文本特征、共词分析、主题分类、情感值计算、数据可视化等，分析论文选题的特征以及论文评语情感倾向。首先，对论文题目进行分词，在分词时需要考虑专业词汇的影响。其次，将文本向量化，即把文本数据变成向量数据。TF-IDF（词频—逆文档频率）是进行文本向量化处理的常用方法，通过计算一个词在文档中的出现频率和在整个语料库中的逆文档频率来评估该词对文档的重要性。最后，采用 LDA 主题模型对论文题目的特征词进行自动分类，即根据给定的主题数量，将论文题目提取的关键词自动划分为设定的主题数量。将 LDA 主题模型与论文题目相结合，最终找出其中较为热门的、能代表专业特点的词汇 [3]。

对于评语的分析，我们需要对论文评语数据进行分词处理和词频统计。评语中包含的词汇如否定词和程度副词，它们的出现直接影响评语的情感倾向，因此对于这些词需要特殊处理。此外，我们将基于情感词典进行情感定位，并建立语料库。情感词典将为我们提供每个词汇的情感值，进而得出整体的评语情感评分。有了情感评分后，我们可以统计正向和负向论文评语的特征。通过这种方式，我们可以深入挖掘评语的情感倾向和评审者的关注维度，最终分析评语的问题表现及成因。

3 研究结果与分析

3.1 基于共词分析法的题目共词网络图谱构建

共词分析法通过对一组词两两统计它们在同一篇文献中出现的次数，并对这些词进行聚类分析。这种方法可以帮助我们了解学科的发展趋势，以及不同时期的研究热点和主题。通过对一组文献的主题词两两在同一篇文献中出现的频率进行统计，可以形成一个由这些词对组成的共词网络，网络内节点之间的远近便可反映主题内容的亲疏关系。在共词网络图中，边的权重通常表示共现的次数，节点代表关键词，节点的大小和颜色等属性也可以反映其在网络中的重要性和其他特征 [4]。论文中，通过共词分析法对论文题目关键词进行网络图谱构建如图 1 所示，可以评估论文选题的创新性、研究热点和方向。

从图 1 中可以看出，学生选题采用 Java、Spring、SpringBoot 技术进行系统开发的频率高于其他，这表明这些技术在学生中具有较高的认可度和应用价值，学生在技术选型时，倾向于选择自己熟悉和掌握的领域，这既是为了保证项目的可行性和效率，也是为了确保开发过程中的稳定性和质量。Java 生态系统因其强大的功能性和广泛的适用性，自然成为学生的首选。同时，学生选题多集中于与学生的学习、校园生活和成长相关的一些课题。这些课题直接与学生个体的日常生活和学习经验紧密相连，因此学生更愿意选择这些与其息息相关的领域进行深入研究和实践。

3.2 基于 LDA 模型的论文主题挖掘

在本次研究中，我们通过 LDA 算法对论文题目进行主题分类，发现论文题目主要涉及以下四个主题：校园生活与学生事务、健康与养老服务、乡村振兴与旅游、本地生活服务。其中，校园生活与学生事务和本地生活服务是最为热门的主题，分别占据了 37% 和 31% 的比例，乡村振兴与旅游占据 17% 的比例，健康与养老服务占 15% 的比例。各主题下的关键词分布情况如表 2 所示。将四个主题的数据进行数据可视化，生成主题分布图如图 2 所示。从表 2 和图 2 中可以看出，在校园生活与学生事务这个主题中，我们可以看到很多关键词都与学生的学习、生活和成长密切相关，这些词代表了学生们在校园生活中所关注的方面，也为学生们的成长提供了方向和指导。随着社会老龄化的加剧，健康与养老服务成为人们关注的热点话题。学生们对于老年人的健康和养老问题非常关注，也反映了他们对于社会公益事业的积极参与。同时，乡村振兴与旅游是当前国家大力发展的战略之一。学生们对于乡村旅游和农业发展非常关注，也反映了他们对于国家战略的积极响应。

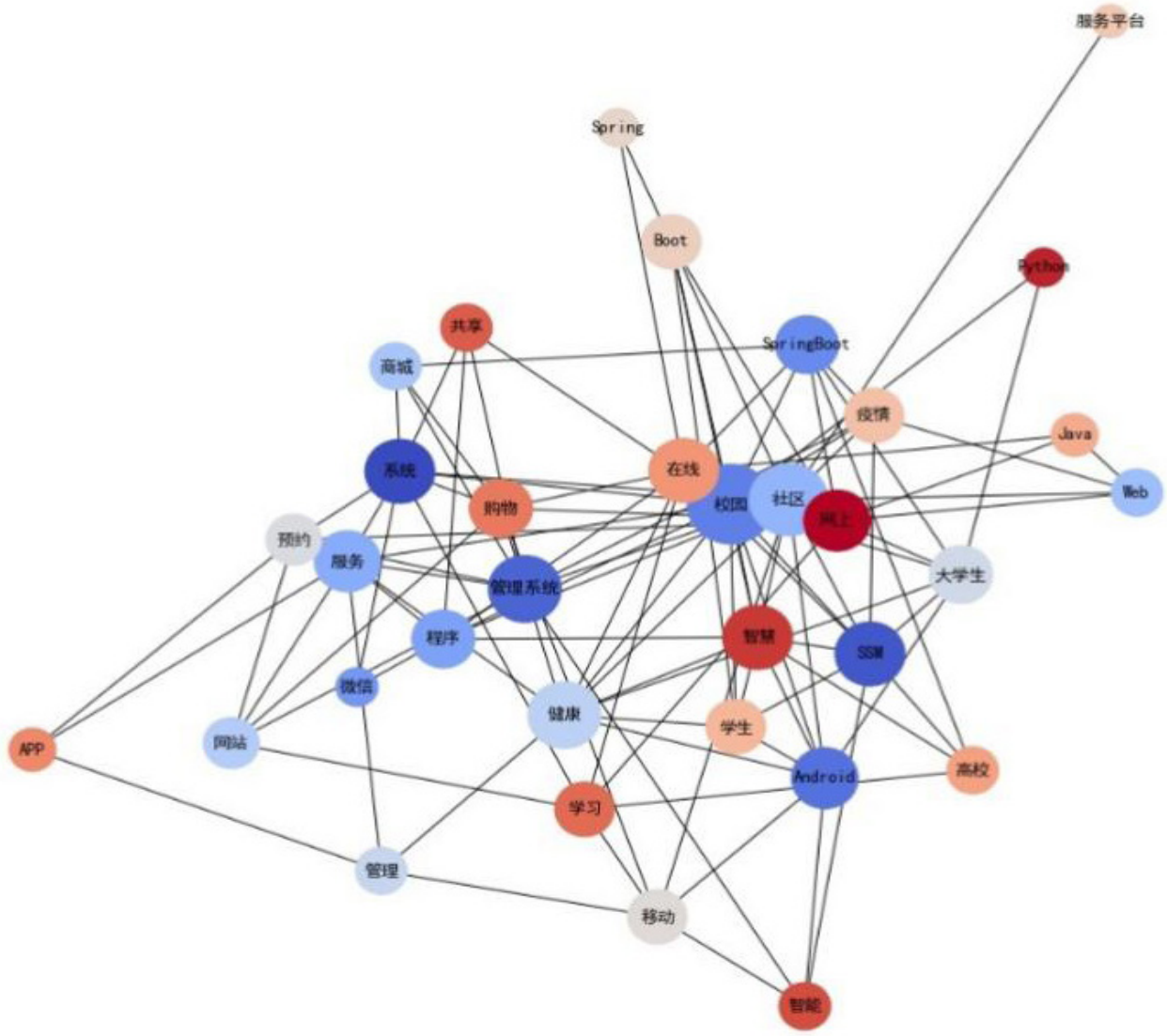

图 1 论文题目共词网络图

表 2 论文主题与关键词分布情况

主题类别	Top20 主题关键词
校园生活与学生事务	校园、大学生、高校、学习、实验室、考试、校友、竞赛、考勤、文化选课、图书馆、就业、社团、图书、考研、教育、教学、课程、实习
健康与养老服务	健康、心理健康、健身、医院、健身房、病历、心理咨询、老年人、敬老院、养老院、养老、志愿者、康养、医养、爱老、购药、门诊、疫苗运动、体育用品
乡村振兴与旅游	旅游、农产品、乡村、酒店、农业、水果、振兴、鲜花、旅游景点、民宿、红色旅游、农村、特产、农电、助农、展销、导游、蔬菜水果、扶贫、公益活动
本地生活服务	社区、公益、点餐、物业、餐饮、订餐、生活、记账、超市、外卖、零食房屋、跑腿、电影、美食、资讯、视频、快递、家政、家庭

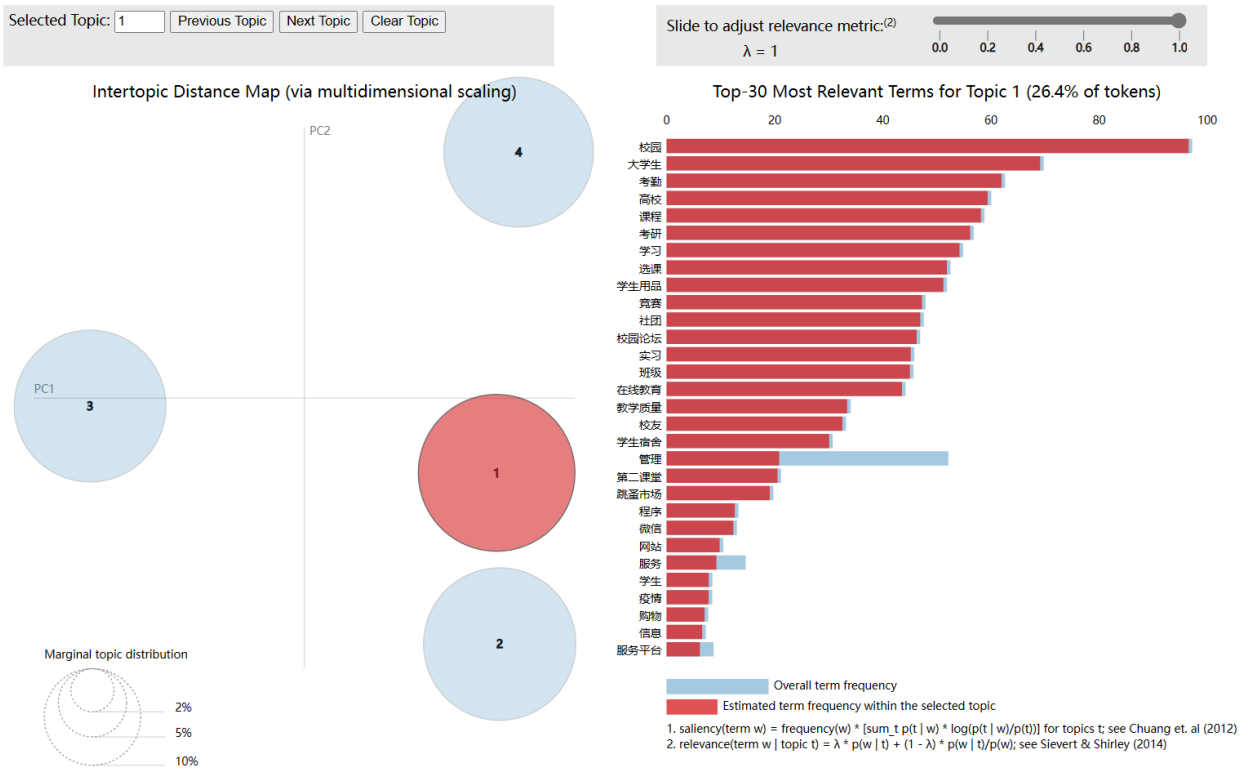

图 2 主题分布离散可视化图

3.3 基于情感值计算的论文评语分析

论文中，利用情感词典对预处理后的文本进行情感词汇识别，提取出文本中的情感词汇进行情感词计算，得出每一条评语的情感倾向为正向还是负向。其中，分别对正向和负向的评语进行主题词挖掘，最终得出专家评审时主要关注的维度，以及学生在论文写作中经常出现的代表性问题。实验结果表明，专家评审意见中主要关注选题意义、专业能力、逻辑构建、学术规范以及工作量这五个维度。在选题方面，评语在提到更多的是该题目是否具有明确的应用价值以及是否具有前瞻性和创新性[5]。在专业能力方面，集中于学生是否具备解决该问题的必要能力，所采用的方法和技术是否合适和先进。在逻辑构建方面，评审专家会考虑论文的引言、方法、结果和讨论等部分是否清晰，各部分之间的逻辑关系是否严密。在学术规范方面，专家会评估论文图表和数据是否清晰、准确。对于工作量，专家会评估论文的篇幅、研究过程的复杂性和实验数据的数量等因素，以判断研究者是否投入足够的时间和精力来完成这项研究。

4 结论和建议

通过本文分析，可以为本科毕业论文指导提供以下几点改进措施和管理建议：需要加强选题指导，在选题阶段给予学生更多的指导，帮助学生选择既符合专业方向又具有实际意义的题目。针对学生专业能力的不足，可以加强相关培训和指导，定期举办学术讲座、研讨会，邀请专家对学生进行专业知识和技能的培训。在论文指导中还应注重培养学生的逻辑思维能力，帮助学生理清论文思路，确保论文结构清晰、逻辑严密。教师应当全面考虑学生的工作量，避免只看论文长度而忽视研究过程的复杂性和工作量。

参考文献

[1] 张磊.民办本科计算机类专业毕业设计选题现状与对策[J].计算机教育,2022(3):6-10.

[2] 邓磊波.理工科本科毕业论文存在的问题及质量提升建议[J].科技视界,2021(26):171-173.

[3] 任福,刘晓宇,石长虹,等.GIS本科毕业设计选题特征分析——以武汉大学为例[J].地理信息世界,2019,26(6):129-132+138.

[4] 李佳晨.基于语义挖掘的论文题目与评语的分析与应用[D].武汉:华中师范大学,2018.

[5] 何中清.系统功能语言学视角下的学术话语分析范式建构[J].外语学刊,2021(2):23-27.