**复旦大学本科生毕业论文（设计）开题报告**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 |  | | 学号 |  |
| 所在院系 | 计算机科学与技术 | | 专业 | 计算机科学与技术 |
| 指导教师 |  | | 职称 |  |
| 校外指导教师  及其所属单位 | 无 | | 职称 | 无 |
| 论文（设计）题目 | 无示例文本分类研究 | | | |
| **开题报告：建议包含以下内容（可另附页）：**  1. 选题的目的和意义；  2. 国内外相关研究状况综述（列出相应的参考文献）；  3. 主要研究内容与基本思路，详细技术路线，并分析可行性、难点和创新点；  4. 预期成果及形式。 | | | | |
| **研究进度及具体时间安排（不够写可加行）** | | | | |
| 起止日期 | | 主要研究内容 | | |
| 2019.03.07-2019.03.21 | | 查阅相关文献，深入了解研究课题的内容 | | |
| 2019.03.21-2019.04.01 | | 选定实验对象及数据，并选择适合的分类方法作为备用策略 | | |
| 2019.04.01-2019.05.01 | | 实现一个改进的无示例文本分类系统，并开始撰写论文 | | |
| 2019.05.01-2019.06.01 | | 完成撰写论文，并准备论文答辩 | | |
| **指导教师对课题报告的意见：**  **1．对选题依据、基本思路或技术路线的可行性、创新性的评价；2．存在的主要问题和改进建议。**  指导教师签名： 年 月 日 | | | | |

**一、选题背景、目的及意义**

随着计算机和通信技术的迅速发展，计算机应用渗入到了人类社会生活的方方面面。大量极具研究和应用价值的数据被生产出来并被存储到计算机系统中，而文本数据正是其中的关键数据之一。文本分类能够根据人们的需求，将海量的文本信息初步划分成不同类别的文本信息，以便于进一步的文本处理，最终精确获取所需的文本信息。文本分类被广泛应用于许多领域，包括情感分析，主题标记，文本索引，垃圾邮件检测和信息检索管理等等。

本课题将对比讨论现有的无示例文本分类方法的优缺点和适用范围，并对无示例文本分类方法提出相关改进，以适应和应用于某些具体文本领域。

无示例文本分类方法避免了有监督和半监督学习对训练数据的严格要求和训练过拟合的问题，拓宽了文本分类的使用范围，增强了文本分类的可行性，对当前没有大量优质的文本数据进行训练的语言文本具有重大意义。

**二、研究状况综述**

文本分类技术是信息检索和文本挖掘的重要基础，其主要任务是在预先给定的类别标记(label)集合下，根据文本内容判定它的类别。20世纪90年代以前，文本分类主要依赖于贝叶斯公式，知识工程和专家系统。在此之后，基于机器学习的文本分类方法逐渐成熟起来，更注重分类器的模型自动挖掘和生成及动态优化能力，在分类效果和灵活性上都比之前基于知识工程和专家系统的文本分类模式有所突破。但是由于近年来移动互联网的爆炸式发展，在互联网中分布传播的海量电子化文本逐渐呈现出种类多样、分布偏斜、关系复杂、更新频繁及标注困难等非结构化特征。为解决有监督和半监督的机器学习文本分类方法遭遇的可扩展性差、语料缺乏及精度降低等问题，产生了无示例文本分类方法[[1]](#endnote-0)。

表示文本片段语义的最简单方法是将其视为单词空间中的向量，也叫做词袋（BOW）表示。但是这种表示只使用文本中的单词，无法表达文本的衍生含义。为了获得更有意义的文本片段的语义解释，Gabrilovich和Markovitch引入了显式语义分析（ESA），并使用维基百科作为其世界知识的来源[[2]](#endnote-1)。Chang等出了一种无示例分类模型，使用维基百科作为世界知识来源，并证明单独的标签名称通常足以诱导分类器[[3]](#endnote-2)。Song和Roth提出的无示例层次文本分类方法由语义相似性步骤和自举步骤这两个步骤组成，并且自举无示例分类与具有数千个标记示例的监督分类相比具有竞争力[[4]](#endnote-3)。Ha-Thuc和Renders提出了基于LDA的无监督方法（OHLDA）[[5]](#endnote-4)。之后，Chen和Xia等提出了描述性LDA（DescLDA）模型，该模型仅使用类别描述词和未标记文档执行无示例文本分类（DLTC），不使用显式语义分析（ESA），不需要大规模精细编译的语义知识库作为外部知识来源[[6]](#endnote-5)。Potthast; Sorg and Cimiano等提出了ESA的推广，跨语言显性语义分析（CLESA）。Song和Upadhyay等提出基于CLESA的跨语言无示例分类，将多种外语语言文档和英语标签嵌入到共享语义空间中[[7]](#endnote-6)。Li和Zheng等提出了类别嵌入模型和分层类别嵌入模型这两个模型来同时学习大规模知识库中的实体和类别在语义空间中的表示[[8]](#endnote-7)。Li和Yang等开发了基于伪标签的无示例朴素贝叶斯（PL-DNB），采用期望最大化（EM）算法以半监督方式训练PL-DNB，以高度可接受的置信度迭代地更新伪标签。实验结果表明，PL-DNB优于使用种子词的现有无示例算法，尤其在不平衡数据集上表现良好[[9]](#endnote-8)。

**三、主要研究内容**

1、无示例文本分类的形成过程及意义；

2、对现有无示例文本分类方法进行简单介绍与比较；

3、提出一种在某方面加以改进的性质更好的无示例文本分类方法；

4、将改进后的无示例文本分类方法应用于某些具体文本领域并评价其分类效果。

**四、可行性、难点及创新点**

1、互联网电子文本呈现出种类多样、分布偏斜、关系复杂以及更新频繁等非结构化特征，在处理上比较困难；

2、需要将待分类文本与标签的语义充分精确地表示成可被计算的数据；

3、创新点在于摒弃训练的过程，构建文本语义的数字化表示与计算。

**五、预期成果及形式**

1、实现一个改进的无示例文本分类系统。

2、完成相应的研究论文。

1. 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[D], 2006. [↑](#endnote-ref-0)
2. Gabrilovich E, Markovitch S. Computing semantic relatedness using wikipedia-based explicit semantic analysis[C]//IJcAI. 2007, 7: 1606-1611. [↑](#endnote-ref-1)
3. Chang M W, Ratinov L A, Roth D, et al. Importance of Semantic Representation: Dataless Classification[C]//AAAI. 2008, 2: 830-835. [↑](#endnote-ref-2)
4. Song Y, Roth D. On dataless hierarchical text classification[C]//Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014. [↑](#endnote-ref-3)
5. Ha-Thuc V, Renders J M. Large-scale hierarchical text classification without labelled data[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 685-694. [↑](#endnote-ref-4)
6. Chen X, Xia Y, Jin P, et al. Dataless text classification with descriptive LDA[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015. [↑](#endnote-ref-5)
7. Song Y, Upadhyay S, Peng H, et al. Cross-Lingual Dataless Classification for Many Languages[C]//IJCAI. 2016: 2901-2907. [↑](#endnote-ref-6)
8. Li Y, Zheng R, Tian T, et al. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification[J]. arXiv preprint arXiv:1607.07956, 2016. [↑](#endnote-ref-7)
9. Li X, Yang B. A Pseudo Label based Dataless Naive Bayes Algorithm for Text Classification with Seed Words[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1908-1917. [↑](#endnote-ref-8)