本科毕业论文

论文题目：　基于 LDA 模型和类别关键词的
弱监督文本分类方法的研究

姓　　名：熊倩　学　　号：15307130228

院　　系：计算机科学技术学院

专　　业：计算机科学与技术

指导教师：朱山风　职　　称：副教授

单　　位：复旦大学

完成日期：2019 年 5 月 26 日

基于 LDA 模型和类别关键词的弱监督文本分类方法的研究

熊倩
学号：15307130228
专业：计算机科学与技术

摘要：机器学习技术在近几年飞速发展，也产生出了许多优质的成果，并不断地落实到各个应用场景之中。在文本分类领域中，监督学习能够在给定大量有标注训练集的情况下，完成特定的学习任务。然而，监督学习对训练集的标签、数据平衡度和数据量的依赖，导致其无法使用在数据信息中占绝大部分的互联网上的散布数据，也无法满足日益增加的各类学习需求。为了解决这一问题，本文提出了基于 LDA 主题模型和类别关键词的弱监督文本分类方法 KWC-LDA。KWC-LDA 由两个子分类器经决策优化后得到：类别关键词优化后的 LDA 模型子分类器和类别关键词直接聚类子分类器。在 LDA 模型子分类器中，我们使用无标注的文档集生成 LDA 主题模型，从而对文档集进行分类。在类别关键词直接聚类子分类器中，我们使用类别关键词对文档集进行直接聚类。最终我使用 KWC-LDA 完成了对新闻数据集的弱监督文本分类。

本文的主要贡献有：1) 考虑使用弱监督学习对文本进行分类，这解除了对训练文本集的标签、数据平衡度和数据量的依赖，对更广泛多样的文本集的分类任务而言具有重大意义；2) 本文对文本的预处理过程进行了独特的改进并对 LDA 主题的生成过程进行优化；3) 使用类别关键词在两个不同层次上对 LDA 模型进行优化和补充，从而使 LDA 模型的分类过程更紧贴类别信息而非文档语义结构。

关键字：文本分类，LDA 主题模型，类别关键词，弱监督学习

Abstract: Machine learning technology has developed rapidly in recent years, and has produced many high-quality results, and has been continuously implemented in various application scenarios. In the field of text categorization, supervised learning can accomplish a specific learning task given a large number of labeled training sets. However, supervised learning relies on the training set's labeling, data balance, and data volume, making it impossible to use the vast majority of the data on the Internet to spread the data, and can not meet the increasing variety of learning needs. In order to solve this problem, this paper proposes a weakly supervised text classification method KWC-LDA based on LDA theme model and category keywords. KWC-LDA is optimized by two sub-classifiers: LDA model sub-classifier and category keywords direct sub-classifier. In the LDA model sub-classifier, we use the unlabeled document set to generate the LDA topic model to classify the document set. In the category keywords direct clustering sub-classifier, we use the category keywords to directly cluster the document set. Eventually I used KWC-LDA to complete the weakly supervised text classification of the news dataset.

The main contributions of this paper are as follows: 1) Consider using weakly supervised learning to classify text, which removes the dependence on the label, data balance and data volume of the training text set, and is of great significance for the classification task of a wider variety of text sets; 2) The text preprocessing process is uniquely improved and the LDA topic generation process is optimized. 3) The LDA model is optimized and supplemented at two different levels using category keywords, thus making the LDA model classification process target at category information instead of document semantic structure.

Keywords: text categorization, latent Dirichlet allocation model, keywords of topics, weakly supervised learning

目录

第一章 绪论

1.1 研究背景和意义

随着互联网的不断普及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重要数据之一。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。

在文本分类领域中，监督学习能够在给定大量有标注训练集的情况下，完成特定的学习任务。然而，监督学习对训练集的标签、数据平衡度和数据量的依赖，导致其无法使用在数据信息中占绝大部分的互联网上的散布数据，也无法满足日益增加的各类学习需求。为了解决这一问题，本文提出了基于 LDA 主题模型和类别关键词的弱监督文本分类方法 KWC-LDA。

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，对当前没有大量且优质的文本数据进行训练的语言文本具有重大意义。

1.2 当前研究状况综述

文本分类技术是信息检索和文本挖掘等领域的重要基础，其主要任务是在预先给定的类别标签(label) 集合下，对文本内容进行处理和分析进而判定当前文本的类别。20 世纪 90 年代以前，文本分类任务主要依赖于贝叶斯公式[1]，知识工程[2]和专家系统[3]等技术。

在此之后，基于机器学习的文本分类方法逐渐成熟起来。相比于之前基于知识工程以及专家系统的文本分类方法，使用机器学习技术来对文本进行分类，得到的分类模型往往具有数据挖掘自动化和参数动态优化的能力，并能够提升分类效果和增加分类方法的灵活性。

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文本越来越呈现出类型多样、分布偏斜、质量低劣、更新频繁及标注困难等非结构化特征。在对互联网文本进行分类时，有监督和半监督的机器学习文本分类方法遭遇了可扩展性差、语料缺乏及随之而来的精度降低等问题。因此随后产生了弱监督或者无示例的文本分类方法。

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[4]。Chang 等提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[5]。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, 实验表明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[6]。Ha-Thuc 和 Renders 提出了基于 LDA 的无监督方法(OHLDA)[7]。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，不需要大规模精细编译的语义知识库作为外部知识来源[8]。Potthast, Sorg 和 Cimiano 等提出了 ESA 的推广, 跨语言显性语义分析 (CLESA)[9]。在此之后，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[10]。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[11]。Li 和 Yang 等开发了基于伪标签的无示例朴素贝叶斯算法(PL-DNB)[12]，该算法采用期望最大化(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。

1.3 研究内容和主要贡献

在仅使用 LDA 模型进行文本分类时，通常包含以下三步：1、在给定的数据集上构建无监督 LDA 模型，并生成一组原始主题; 2、手动将每个主题映射到一个类别标签上，然后将映射到同一标签的各个主题组合到一个新的主题中; 3、使用无监督 LDA 模型学习到的文档-主题后验概率对测试文档进行分类。然而，LDA 模型的生成仅仅使用了文档集的语义结构，而文档集的语义结构可能会与实际的类别标签有所偏差，导致生成的主题只是在描述语义结构而非类别。所以我们考虑使用类别信息对 LDA 模型的主题生成过程进行改进，从而更好地表示文档集的类别。

本文提出的基于 LDA 模型和类别关键词的弱监督文本分类方法 KWC-LDA，使用了类别关键词在两个不同层次上对 LDA 模型进行改进: 一是基于 LDA 模型分类文本时使用类别关键词对训练文本集进行类别信息的扩充与优化; 二是使用类别关键词对文本集进行直接聚类从而对文本进行分类。最终的文本分类方法由

上述类别关键词优化后的 LDA 模型子分类器和类别关键词直接聚类子分类器综合得到，综合子分类器的方式为决策优化，即最终的分类结果是由两个子分类器的分类结果线形加权得到的。

本文的贡献主要体现在以下几个方面：

1）本文考虑使用弱监督学习对文本进行分类，这解除了对训练文本集的标签、数据平衡度和数据量的依赖，对更广泛多样的文本集的分类任务而言具有重大意义。

2）在文本的预处理过程中进行了独特的改进并对 LDA 主题的生成过程进行优化。

3）使用了类别关键词在两个不同层次上对 LDA 模型进行优化和补充，从而使 LDA 模型的分类过程更紧贴类别信息而非文档语义结构。

第二章 相关原理介绍

2.1 弱监督文本分类

弱监督文本分类，或者叫无示例文本分类(dataless text classification)，并不需要有标注的文档集作为训练数据，能够对输入的无标注文档集直接进行分类。弱监督文本分类方法可以分为两种类型：基于分类和基于聚类。基于分类的方法一般采用自动算法来创建机器标记的数据。但是基于分类的方法会存在机器标记数据的质量难以控制的问题，从而可能导致不可预测的偏差。相比之下，基于聚类的方法首先使用基于类别标签或描述的模型，来计算文档之间的相似性，并对测试文档进行聚类，最后将聚类分配到类别。 此外，无示例文本分类器能够通过理解标签来准确地对文本文档进行分类，或者基于主题模型将文档划分成不同的主题以进行文本分类。

2.2 主题模型

在自然语言理解任务中，我们可以通过一系列的层次来提取含义：从单词、句子、段落，再到文档。而在文档层面，理解文本最有效的方式之一就是分析其主题。在文档集合中学习、识别和提取这些主题的过程被称为主题建模。主题模型基于的基本假设是：每篇文档由多个主题所构成，同时每个主题由多个单词所构成。例如，在**图 1** 中，这里存在两篇文档，同时每篇文档分别包含三个不同的主题。每个主题会关联属于该主题的单词，而每个单词则会属于一个或多个主题。

文档	主题	单词
最先进的集成电路是微处理器或多	集成电路	**集成电路**、处理器、计算机
核处理器的核心，可以控制计算机	处理器	集成电路、**处理器**、计算机
到手机到数字微波炉的一切。	计算机	集成电路、处理器、**计算机**
除了商业主流的操作系统外，从	商业	**商业**、操作系统、开源
1980 年代起在开源代码的世界中，	操作系统	商业、**操作系统**、开源
BSD 系统也发展了非常久的一段	开源	商业、操作系统、**开源**
时间。		

图 1：一个主题模型的例子

实际上，文档的语义由一些我们所忽视的隐变量或潜变量管理。因此，主题建模的目标就是揭示这些潜在变量，也就是主题，正是它们塑造了我们文档和语料库的含义。当下最流行的主题建模技术有潜在语义分析(latent semantic analysis, LSA)[17]，Probabilistic LSA(PLSA)[18] 和潜在狄利克雷分配(latent Dirichlet allocation, LDA)[19]。

潜在语义分析(LSA) 的主要思想是把现有的文档-术语矩阵分解成相互独立的文档-主题矩阵和主题-术语矩阵。LSA 方法快速简便，但它的缺点是：缺乏可解释性，依赖大量的输入文本来获得准确的结果，同时表征效率比较低。概率潜在语义分析技术(pLSA) 则采用概率方法替代 LSA 方法中的 SVD 技术以解决问题。pLSA 的主要思想是找到一个潜在主题的概率模型，该模型可以生成我们在文档-术语矩阵中观察到的数据。但 pLSA 的参数数量随文档数线形增长，容易出现过拟合的问题。LDA 主题模型即潜在狄利克雷分配，是 pLSA 的贝叶斯版本。它使用狄利克雷先验来处理文档-主题和主题-单词分布。一般来说，LDA 比 pLSA 效果更好，因为它可以更容易地泛化到新文档中去。

第三章 LDA 模型与类别关键词

3.1 基于 LDA 模型的文本表示

3.1.1LDA 模型的三层结构

LDA 主题模型是一个基于三层贝叶斯概率运算的文本主题生成模型, 使用贝叶斯概率将文档，主题和单词三层结构联系起来。具体而言，一个 LDA 模型可能是由足球主题和篮球主题共同构成的。同时，每个主题生成词汇表中的各单词具有不同的概率。例如足球主题能生成足球、草场和射门等单词，并且在足球主题下，生成足球单词的概率本身会很高。同样地，篮球主题同样具有生成各个单词的概率：其中，篮球，二分和三分可能具有较高的概率。而与各主题没有特殊相关性的单词，例如人们，在各个主题中的生成概率大致相同。在 LDA 模型中，生成的主题是在自动检测单词共现可能性的基础上识别的, 而非语义上的或者是认识论上强制定义的。

它以输入的无标注文本集作为训练数据，根据文本集中的潜在主题和单词的分布结构，生成文档到主题的概率分布和主题到单词的概率分布。借助文档到主题的概率分布以及主题与类别的对应关系，我们能够对文档进行无监督分类。但是在原有的 LDA 模型中，主题的生成完全基于文档的单词信息，并未考虑文档中所蕴含的围绕类别的信息, 所以生成的文档主题结构跟文档集的类别会有所出入。这里使用概率图形模型(PGM) 中常用的平板表示法，来说明 LDA 模型中各个变量之间的依赖关系，如**图 2** 所示：

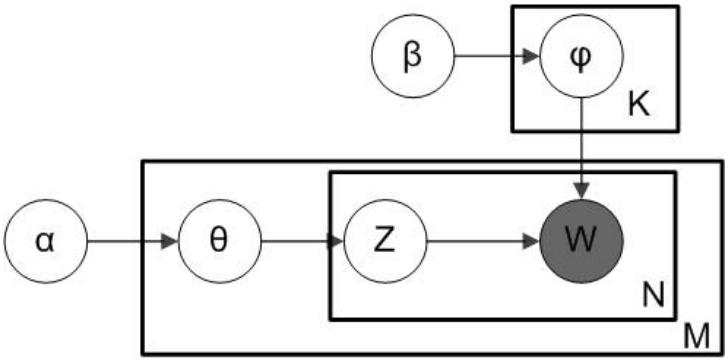

图 2：描述 LDA 模型的平板表示图

图 2 中的框是代表重复的“板”，它们是重复的实体。下部中的外板表示文

档，而下部里的内板表示给定文档中的重复单词位置，每个位置与主题和单词的选择相关联。M 表示文档的数量，N 表示文档中的单词数。变量名具体含义如下：α是每篇文档关于主题的狄利克雷分布的参数；β是每个主题关于单词的狄利克雷分布的参数；θ_d 是文档 d 的主题概率分布；K 表示主题的总个数；φ_k 表示主题 k 的单词概率分布；$z_{i,j}$ 表示文档 i 中的第 j 个单词的主题；$w_{i,j}$ 表示文档 i 中的第 j 个单词。其中，单词 W 是灰色的，表示变量 $w_{i,j}$ 是可见的，可被观察到的，而图中的其他变量则是隐变量，不能被直接观察到。

3.1.2LDA Gibbs 抽样算法

在 LDA 模型中，由文档 d 组成的文档集 D 的主题模型生成过程如下：

1) 对文档集 D 中的每个文档 d，其主题分布θ_d是 Dirichlet(α)分布的一个抽样
2) 对文档 d 中的每个单词 w:
   (a) 在样本主题分布θ_d中随机抽取一个主题 $z_{i,j}$;
   (b) 在主题 $z_{i,j}$ 的多项式分布中随机抽取一个单词 $w_{i,j}$。

与此同时，我们希望获得文档集中每篇文档属于各个主题的概率（即文档-主题分布）和文档集的词汇表中任一单词在不同主题中所占的比重（即主题-单词分布），从而计算出目前的单词所属的概率最大的主题，并更新该词的主题。目前对 LDA 模型抽样的方法有变分法、期望最大化(Expectation Maximization, EM) [14]及 Gibbs 抽样[15]方法。其中变分法得到的模型与真实情况有所偏差，而 EM 算法往往容易得到似然函数的局部最优解，但很难找到全局最优解。Gibbs 抽样是基于马尔科夫链的蒙特卡洛方法(Markov chain Monte Carlo，MCMC)[16]，它容易实现，并能有效地从输入文本集中抽取主题。因此，这里使用 Gibbs 抽样算法对 LDA 模型进行抽样。

在使用 Gibbs 抽样算法对 LDA 模型进行抽样的过程中，首先需要针对给定的训练文档训练模型，然后根据训练出的模型对测试文档进行分类。LDA Gibbs 抽样算法的训练过程如下：

1）设输入的训练文档中的文档数量为 M，训练文档集合的单词数为 W，并选择合适的主题数 K，选择合适的参数α和β。
2）初始化单词的所属主题：对文档集中每篇文档中的每个单词，随机赋予其一个主题为 z。
3）重新遍历文档集，对其中的每个单词，使用 Gibbs 采样公式更新它的主题，并更新文档集中该词的编号。
4）重复第 3 步基于坐标轴轮换的 Gibbs 采样，直到 Gibbs 采样收敛。

5）统计文档集中的各个文档各个单词的主题，并得到文档-主题分布θ_d；统计文档集中各个主题的单词分布，得到 LDA 模型的主题-单词分布φ_k。

根据 Gibbs 抽样算法[]，单词 z_i 属于主题 j 的后验概率的计算公式如下：

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto P(w_i|z_i = j,\mathbf{z}_{-i},\mathbf{w}_{-i})P(z_i = j|\mathbf{z}_{-i}) \tag{1}$$

其中，**w** 表示文档集中所有的单词，$\mathbf{w}_{-i}$ 表示除了单词 w_i 之外的其它单词，$\mathbf{z}_{-i}$ 表示除了单词 w_i 之外其它单词的主题。

同时，式子右部的第一个概率和第二个概率的最终表达式为：

$$P(w_i|z_i = j,\mathbf{z}_{-i},\mathbf{w}_{-i}) = \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \tag{2}$$

$$P(z_i = j|\mathbf{z}_{-i}) = \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha} \tag{3}$$

其中，W 表示文档集合的单词数，K 表示主题数，α和β是概率分布的参数，$n_{-i,j}^{(w_i)}$ 表示在整个文档集中，单词 w_i 除了当前单词之外，分配给主题 j 的数量；$n_{-i,j}^{(\cdot)}$ 表示在整个文档集中，除了当前单词之外的所有单词分配给主题 j 的数量；$n_{-i,j}^{(d_i)}$ 表示在文档 d_i 中，除了当前单词之外的所有单词分配给主题 j 的数量；$n_{-i,\cdot}^{(d_i)}$ 表示在文档 d_i 中，除了当前单词之外的所有单词分配给任意主题的数量。

因此，我们能够得到单词 z_i 属于主题 j 的后验概率的相关表达式，从而在 Gibbs 抽样算法的步骤 3 中更新单词的主题，其表达式为：

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha} \tag{4}$$

在步骤 5 中，根据文档集中的各个文档中各个单词的主题，计算文档-主题分布θ_d和主题-单词分布φ_k的公式如下：

$$\theta_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{(d)} + K\alpha}, \varphi_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta} \tag{5}$$

其中，$n_j^{(d)}$ 表示文档 d 中属于主题 j 的所有单词的个数；$n_{\cdot}^{(d)}$ 表示文档 d 中属于任意主题的单词总个数；$n_j^{(w)}$ 表示文档集中单词 w 分配给主题 j 的次数；$n_j^{(\cdot)}$ 表示文档集中分配给主题 j 的单词的总个数。公式(5)的含义为：文档 d 属于主题 j 的概率，跟文档中属于主题 j 的单词数与文档中分配主题的单词总数成正比；文档集合中，单词 w 属于主题 j 的概率，跟文档集合中单词 w 属于主题 j 的次数与文档集合中属于主题 j 的单词总数成正比。

3.2 文档集中的类别关键词信息

类别关键词能够被组织起来，共同刻画文本集的类别的语义，是被刻画的文本类别的核心概念。并且一个单词只能成为一个类别的关键词，不能同时作为多个类别的关键词。例如，足球，草场和射门等单词可以是类别足球的类别关键词。草场和射门辅助类别足球自身来详细和深入地刻画类别足球，同时草场和射门不能同时作为其他类别如篮球、羽毛球和游泳的类别关键词。

对于暂时没有类别关键词信息的文本集：我们可以抽取文本集中的有意义的高频词汇作为核心词汇；同时可以使用 LDA 模型生成文本集的主题模型，然后抽取各个主题中的重要单词作为核心词汇；也可以在文档集中找出跟核心词汇语义相近的单词作为核心词汇，如两个单词在语义向量空间中的距离相近，或者两者在文档集中经常同时出现；最后手动从核心词汇中筛选出合适的单词作为各个类别的关键词。

对于已存在类别关键词信息的文本集，我们可以直接使用关键词信息；或者在已有类别关键词信息的基础上，根据上述方法重新抽取核心词汇，然后从核心词汇中筛选出合适的类别关键词。例如对 20Newsgroups 数据集，这里使用了 Song and Roth 提出的类别关键词信息，具体如**表 1** 所示：

表 1：20Newsgroups 数据集中的类别关键词

Label	Keywords of this label
talk.politics.guns	gun fbi guns weapon compound
talk.politics.mideast	israel arab jews jewish muslim
talk.politics.misc	gay homosexual sexual
alt.atheism	atheist christian atheism god islamic
soc.religion.christian	christian god christ church bible jesus
talk.religion.misc	christian morality jesus god religion horus
comp.sys.ibm.pc.hardware	bus pc motherboard bios board computer dos
comp.sys.mac.hardware	mac apple powerbook
comp.graphics	graphics image gif animation tiff
comp.windows.x	window motif xterm sun windows
comp.os.ms.windows. misc	windows dos microsoft ms driver drivers card printer
rec.autos	car ford auto toyota honda nissan bmw
rec.motorcycles	bike motorcycle yamaha
rec.sport.baseball	baseball ball hitter
rec.sport.hocke	hockey wings espn
sci.electronics	circuit electronics radio signal battery
sci.crypt	encryption key crypto algorithm security
sci.med	doctor medical disease medicine patient
sci.space	space orbit moon earth sky solar
misc.forsale	sale offer shipping forsale sell price brand obo

第四章 KWC-LDA 弱监督文本分类方法

4.1 基于 LDA 模型的文本分类方法

4.1.1 类别关键词对文本集的优化

在基于 LDA 模型的子分类器 1 中，为了在文档单词中突出类别关键词的作用，我们使用类别关键词对训练文本集进行类别信息的扩充与优化。其主要步骤为：首先设定在每篇文档中，类别关键词的固定比重阈值 α；接着在文档中寻找各个类别的关键词，如果没有找到任何的类别关键词，则无法对该文档进行类别关键词的优化，不执行任何操作；如果找到类别关键词，那么查看类别关键词的数量占该篇文档单词总数是否超过阈值 α，若超过则不对类别关键词进行扩增，若未超过，则将类别关键词的数目进行倍增，直到超过阈值 α。

在对训练文本集进行类别信息的扩充之后，将扩充后的文本集作为训练数据输入 LDA 模型，并进一步生成主题及其关键词。

4.1.2 无意义主题的吸收和类别分派

在 LDA 模型中，需要设置生成主题的个数。一个显而易见的事实是：生成主题的个数越多，主题就越能够精确地表示文本集的语义结构，同时构建模型的时间也会越长，将主题映射到类别的过程也就越复杂。这也给完善 LDA 模型提供了一个方向：即尽可能生成更多的主题，从而更精确地刻画文本集的语义结构。从可行性和实用性上考虑，在本文中我们将主题的个数设置为能基本表示训练文本集类别的最小个数，从而避免无谓的开销，并将注意力集中在类别关键词的使用上。具体而言，使用 LDA 模型生成最小个数主题的过程如下：

1.设置 LDA 模型生成主题的个数为 2；

2.人工判断生成的主题是否能准确表示训练文本集的类别：若能，则生成主题过程结束；否则，将生成主题的个数加 1，并重复本步骤。

例如，在 politics vs religion 分类任务中，如**表 2** 和**表 3** 所示，设置 LDA 模型的生成主题的个数为 2 时，主题 0 中单词 god 应属于 religion 类别，同时单词 gun 应属于 politics 类别。这就在主题 0 的内部产生了冲突，进而无法把主题 0 分派给 politics 类别或者 religion 类别。所以，将生成主题的个数加 1，设置主题个数为 3。此时，生成的主题及其相关词如表所示。这时三个主题内部的关键词

都不产生冲突，所以能够将主题向类别进行分派：将主题 0 分派给 politics 类别，将主题 1 分派给 politics 类别，将主题 2 分派给 religion 类别。同样地，在 comp vs politics 分类任务中，如**表 4** 和**表 5** 所示，初始设置模型的生成主题的个数为 2 时，生成的主题及其相关词如表所示。主题 0 中单词 ax 是噪音词汇，没有意义，同时其余单词应属于 politics 类别。这表明 2 个主题无法将噪音或者无意义词汇进行吸收，进而无法把主题 0 分派给 comp 类别或者 religion 类别，需要增加主题个数以吸收噪音词汇或无意义词汇。所以，将生成主题的个数加 1，设置主题个数为 3。此时，生成的主题及其相关词如表所示。这时，主题 1 和主题 2 内部的关键词都不产生冲突，而主题 0 则吸收掉主要的噪音词汇。所以我们能够将主题向类别进行分派，分派结果为：主题 0 不分派给任何类别，将主题 1 分派给 comp 类别，将主题 2 分派给 politics 类别。同样地，对其他分类任务，我们能够根据上述主题生成算法，来生成能准确表示训练文本集的类别的最小数目的主题。

在将 LDA 模型生成的主题向类别进行分派之后，根据每篇文档从属于各主题的概率，能够进一步得到每篇文档从属于各类别的概率，从而得到类别关键词优化后的 LDA 模型子分类器(LDA model sub-classifier)。

表 2：在 politics vs religion 分类任务中，设置主题个数为 2 时的生成结果

主题	主题关键词
0	0.008*"would" + 0.007*"people" + 0.007*"god" + 0.007*"say" + 0.005*"think" + 0.005*"know" + 0.005*"make" + 0.005*"gun" + 0.004*"get" + 0.004*"go"
1	0.007*"armenian" + 0.007*"say" + 0.007*"people" + 0.007*"israel" + 0.005*"israeli" + 0.005*"go" + 0.004*"arab" + 0.004*"state" + 0.004*"jew" + 0.004*"turkish"

表 3：在 politics vs religion 分类任务中，设置主题个数为 3 时的生成结果

主题	主题关键词
0	0.008*"people" + 0.008*"would" + 0.008*"gun" + 0.006*"make" + 0.006*"get" + 0.005*"say" + 0.005*"state" + 0.005*"go" + 0.005*"government" + 0.005*"think"
1	0.009*"armenian" + 0.008*"israel" + 0.007*"say" + 0.007*"israeli" + 0.007*"people" + 0.005*"arab" + 0.005*"jew" + 0.005*"turkish" + 0.004*"war" + 0.004*"go"
2	0.014*"god" + 0.008*"say" + 0.008*"christian" + 0.007*"would" + 0.006*"people" + 0.006*"think" + 0.006*"know" + 0.005*"believe" + 0.005*"may" + 0.004*"make"

表 4：在 comp vs politics 分类任务中，设置主题个数为 2 时的生成结果

主题	主题关键词
0	0.059*"ax" + 0.008*"people" + 0.007*"say" + 0.006*"go" + 0.005*"would" + 0.005*"turkish" + 0.004*"government" + 0.004*"greek" + 0.004*"think" + 0.004*"party"
1	0.007*"get" + 0.007*"window" + 0.007*"file" + 0.006*"problem" + 0.005*"know" + 0.005*"system" + 0.004*"would" + 0.004*"work" + 0.004*"program" + 0.004*"drive"

表 5：在 comp vs politics 分类任务中，设置主题个数为 3 时的生成结果

主题	主题关键词
0	0.354*"ax" + 0.023*"max" + 0.004*"wm" + 0.004*"bxn" + 0.003*"tm" + 0.003*"giz" + 0.003*"wt" + 0.003*"cx" + 0.003*"ww" + 0.002*"mv"
1	0.009*"window" + 0.009*"file" + 0.006*"get" + 0.006*"problem" + 0.005*"program" + 0.005*"system" + 0.005*"drive" + 0.005*"know" + 0.004*"thank" + 0.004*"work"
2	0.008*"people" + 0.007*"say" + 0.006*"go" + 0.006*"would" + 0.005*"think" + 0.005*"get" + 0.004*"turkish" + 0.004*"government" + 0.004*"see" + 0.004*"right"

4.2 类别关键词对文本集的直接聚类

在使用 LDA 模型的生成主题对文档集进行分类时，我们有时会遇到文档中的关键词无法将该文档归入关键词所属主题的情况。例如，一篇类别为 religion 的文档，可能其中只出现很少数量的 religion 类别关键词，而出现大量的普通词汇如 people，say，go，would，think 等单词。但是这些单词却在 politics 类别的文档中出现的次数更多，因此这篇文档中的普通词汇将这篇文档归入到 politics 类别中。

为解决上述普通词汇干扰类别关键词的问题，我们考虑使用文档中的类别关键词对文档集进行直接聚类。使用类别关键词进行聚类主要是根据文档中各个类别的关键词的数目占比，其具体过程如下：首先在当前文档中查找各个类别的关键词，如果类别关键词数目为 0 则将该文档属于各个类别的概率设为相等，如果查找到各个类别的关键词，则分别将该文档属于各个类别的概率设为该类别的关键词数目占总关键词数目的比值，并由此得到类别关键词直接聚类子分类器(keywords sub-classifier)。

4.3 结合 LDA 模型和类别关键词的 KWC-LDA 方法

最终的 KWC-LDA(latent Dirichlet alocation with keywords of classes)文本分类方法由上述类别关键词优化后的 LDA 模型子分类器和类别关键词直接聚类子分类器综合得到，综合子分类器的方式为决策优化，即最终的分类结果是由两个子分类器的分类结果线形加权得到的。

第五章 实验与分析

5.1 实验设置

KWC-LDA 文本分类算法的运行在 macOS 操作系统上, 使用 Python 语言进行编程，借助 gensim 库实现 LDA 模型。这里使用了开源的 20Newsgroups 以及 Reuters 新闻数据集进行实验。

20Newsgroups 数据集是由近 20000 篇文档组成，被划分为 20 个小类别和 6 个大类别的新闻数据集。该数据集中的训练数据和测试数据分别占比 60%和 40%。20Newsgroups 数据集的类别标签并不是一个完整的单词，而是例如“comp.sys.ibm.pc.hardware”这样的连续符号。Chang 等根据该数据集中的每个类别标签所对应的文档集, 自动地将该类别标签扩展成几个完整的单词, 扩展出的单词就是类别关键词。在此之后，Song 和 Roth 进一步对类别关键词进行增加和完善。本文使用了 Song 和 Roth 提出的类别关键词信息。

另外使用到的 Reuters 新闻数据集来自于 Python 的 keras 库，其中包含来自路透社的 11228 条新闻，总共被分为 46 个主题。Reuters 数据集是发布在 1986 年由一系列短新闻及对应话题组成的数据集, 它是文本分类问题中常用的小数据集，所以本文也使用到该数据集。

5.2 性能度量

精确率是指计算我们预测出来的某类样本中, 有多少是被正确预测的, 这是针对预测样本来说都。而召回率是指针对原先实际样本而言, 有多少样本被正确的预测出来了。F1 分数是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和召回率, 可以看作是模型精确率和召回率的一种加权平均。Macro-F1 则计算出每一个类的精确率和召回率后计算 F1，最后将 F1 平均。

这里用实验结果的 Macro-F1 来作为算法的性能度量标准。同时对于每组实验，重复做十次，然后计算它们平均的 Macro-F1 值。

5.3 文本分类结果对比

5.3.1 类别关键词对文档识别率的提高

这里举例说明在 LDA 模型中，使用类别关键词对输入的文档集进行类别信息上的优化能够提升包含较多常见单词的文档的识别率。

在 politics vs religion 分类任务中，设置主题数为 5 时，有较好的分类效果，Macro-F1 值达到 0.873。但是在生成的主题中，只有一个主题属于 religion 类，其余四个主题都属于 politics 类，同时 politics 类会包含较多的常用词，所以文档的分类容易朝 politics 类倾斜。在 165 个错误分类结果中，有 156 个结果是将 religion 类分类为 politics 类。对这 156 个 religion 类的单词，随着增加类别关键词次数的增长，这些单词属于 religion 类的概率得到矫正与增长。

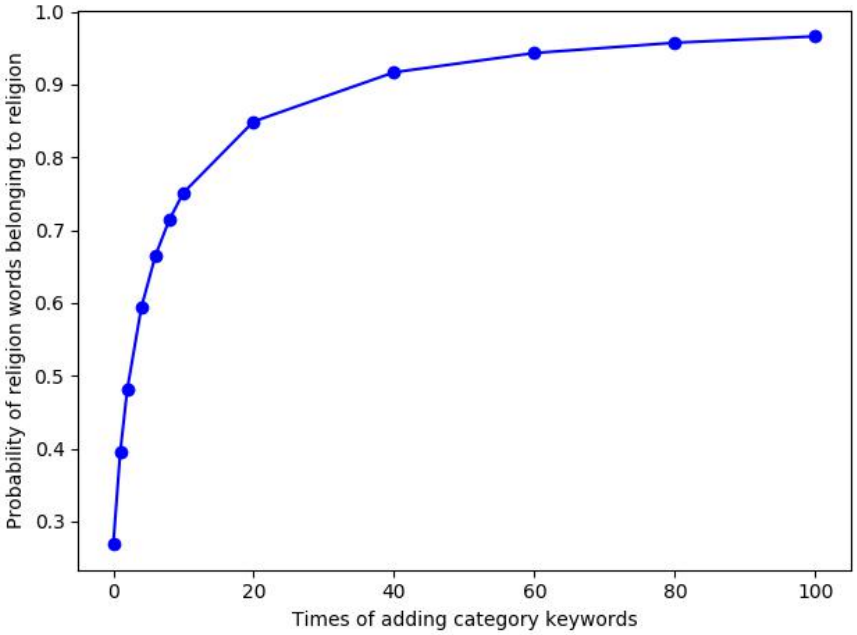

图 3：politics vs religion 分类任务中，类别关键词对文档识别率的提高示例

5.3.2 不同方法的分类结果对比

为评估本文提出的改进算法的分类性能，这里对 KWC-LDA 方法，Clasify-LDA 算法以及 LDA 子分类器和类别关键词聚类子分类器执行同一分类任务的运行结果进行比较。运行结果如下表所示：

由表可知，在以下的分类任务中，LDA 子分类器的性能与 Clasify-LDA 算法相比略有改善。基于类别关键词的子分类器则勉强能对使用的文档集进行分类。

而将 LDA 子分类器和类别关键词聚类子分类器相结合的 KWC-LDA 方法在分类能力上得到进一步的提高。

表 5：KWC-LDA 方法，Clasify-LDA 算法以及两个子分类器执行相同任务的分类结果对比

分类任务	Classify-LDA	KWC-LDA	LDA sub-classifier	keywords sub-classifier
20Newsgroup				
comp vs politics	0.963	0.981	0.969	0.735
comp vs religion	0.955	0.963	0.958	0.679
politics vs religion	0.873	0.891	0.877	0.683
comp vs politics vs rec	0.937	0.945	0.941	0.702
comp vs rec vs religion	0.938	0.942	0.938	0.627
comp vs politics vs rec vs religion	0.892	0.908	0.899	0.530
Reuter	0.831	0.850	0.836	0.686

第六章 总结与改进

6.1 总结

在更广泛的文本分类应用场景下, 无监督或者弱监督的学习方法比有监督的方法更经济实用。本文提出了基于 LDA 模型与类别关键词的弱监督的文本分类算法 KWC-LDA，该算法首先对 LDA 模型的训练文本集进行类别信息的扩充与优化，从而使生成的主题与类别的关系更加紧密。然后再生成能准确表示训练文本集的类别的最小数目的主题，以得到一个文档集的分类结果，避免了过量主题的生成。另外，本文还使用了类别关键词来直接对文档集进行聚类，从而得到另一个文档集的分类结果。最后的分类结果由以上两个分类结果进行决策优化后得到。本方法与 Classify-LDA 以及两个子分类方法在相同分类任务的运行结果，验证了 KWC-LDA 方法的有效性。

6.2 改进

未来可进一步改进的地方有：

1）考虑构建更完善的方法框架来使用文档集的浅层语义信息，如高频词汇，类别关键词，文档中的共现词汇等对文档集进行分类。

2）对使用方法和参数结构进行优化，设置更少和更准确的参数来达到更好的文本分类效果。

3）考虑构建更加自动化的弱监督文本分类方法的分类流程。

致谢

时光荏苒，日月如梭，不觉进入大学已经四年。四年很长，其间丰富的学习与经历令我收获无数；四年又很短，在一日又一日间倏忽掠过，不见踪影。而在这最后的时刻，我想向在此期间给予我鼓励、支持和帮助的人表示以真心的感谢和祝福!

首先，要感谢的是我的导师朱山风老师。本文正是在朱山风老师的精心指导下完成的。在论文的写作过程中，我遇到挫折和困难，老师总是耐心积极地帮助我解决问题，并指出后面学习知识和完善论文的方向。朱老师高尚的人格、敬业的精神、务实的作风，深厚的学术涵养，深深地感染和激励着我。能够得到朱老师的言传身教，我感到非常幸运。其次，要感谢各学长的提点和帮助以及室友在专业技术和学习规划等问题上的合作、支持与帮助。最后，感谢我的父母，父母之心，深沉诚切，欲报之德，昊天罔极。父母的坚定支持和无言奉献一直默默陪伴着我，希望自己能够成为他们的骄傲，也希望能够尽自己的力量去报答和回馈他们。

再次向所有给予我帮助和关心的老师、同学和家人表示衷心地感谢!

参考文献

[1] Bernardo, José M., and Adrian FM Smith. Bayesian theory. Vol. 405. John Wiley & Sons, 2009.

[2] Studer, Rudi, V. Richard Benjamins, and Dieter Fensel. "Knowledge engineering: principles and methods." Data & knowledge engineering 25, no. 1-2 (1998): 161-197.

[3] HayesRoth, Frederick, Donald A. Waterman, and Douglas B. Lenat. "Building expert system." (1983).

[4] Gabrilovich, Evgeniy, and Shaul Markovitch. "Computing semantic relatedness using wikipedia-based explicit semantic analysis." In IJcAI, vol.7, pp. 1606-1611. 2007.

[5] Chang, MingWei, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. "Importance of Semantic Representation: Dataless Classification." In Aaai, vol. 2, pp. 830-835. 2008.

[6] Song, Yangqiu, and Dan Roth. "On dataless hierarchical text classification." In Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

[7] Ha-Thuc, Viet, and Jean-Michel Renders. "Large-scale hierarchical text classification without labelled data." In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 685-694. ACM, 2011.

[8] Chen, Xingyuan, Yunqing Xia, Peng Jin, and John Carroll. "Dataless text classification with descriptive LDA." In Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[9] Sorg, Philipp, and Philipp Cimiano. "Exploiting Wikipedia for cross-lingual and multilingual information retrieval." Data & Knowledge Engineering 74 (2012): 26-45.

[10] Song, Yangqiu, Shyam Upadhyay, Haoruo Peng, and Dan Roth. "Cross-Lingual Dataless Classification for Many Languages." In IJCAI, pp. 2901-2907. 2016.

[11] Li, Yuezhang, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia Sycara. "Joint embedding of hierarchical categories and entities for concept categorization and dataless classification." arXiv preprint arXiv:1607.07956 (2016).

[12] Li, Ximing, and Bo Yang. "A Pseudo Label based Dataless Naive Bayes Algorithm for Text Classification with Seed Words." In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1908-1917. 2018.

[13] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[14] Moon, Todd K. "The expectation-maximization algorithm." IEEE Signal processing magazine 13, no. 6 (1996): 47-60.

[15] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101, no. suppl 1 (2004): 5228-5235.

[16] Gilks, Walter R., Sylvia Richardson, and David Spiegelhalter. Markov chain Monte Carlo in practice. Chapman and Hall/CRC, 1995.

[17] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." Discourse processes 25, no. 2-3 (1998): 259-284.

[18] Hofmann, Thomas. "Probabilistic latent semantic analysis." In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 289-296. Morgan Kaufmann Publishers Inc., 1999.

[19] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.