中图分类号：TP301
论 文 编 号：10006BY1206142

北京航空航天大學

博 士 学 位 论 文

基于主题模型的短文本分析

作 者 姓 名　左源
学 科 专 业　计算机软件与理论
指 导 教 师　许可　教授
培 养 院 系　计算机学院

Topic Model Based Short Text Analysis

A Dissertation Submitted for the Degree of Doctor of Philosophy

Candidate: Yuan Zuo

Supervisor: Prof. Ke Xu

School of Computer Science and Engineering

Beihang University, Beijing, China

中图分类号: **TP301**
论 文 编 号: **10006BY1206142**

博 士 学 位 论 文

基于主题模型的短文本分析

作 者 姓 名 左源 申请学位级别 工学博士
指导教师姓名 许可 张辉 职 称 教授
学 科 专 业 计算机软件与理论 研 究 方 向 文本挖掘
学习时间自 2012 年 09 月 01 日 起 至 2017 年 07 月 01 日止
论文提交日期 2017 年 04 月 25 日 论文答辩日期 2017 年 05 月 30 日
学位授予单位 北京航空航天大学 学位授予日期 2017 年 月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名：_____ 日期： 年 月 日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名：_____ 日期： 年 月 日

指导教师签名：_____ 日期： 年 月 日

摘　要

随着各式各样网络应用的兴起，特别是诸如 Twitter 和 Facebook 等在线社交网络的蓬勃发展，短文本已经成为互联网上信息的主要表现形式。例如，Twitter 上每天大约有 3.19 亿活跃用户，他们能够产生 5 亿左右的 tweets。海量的短文本中蕴含着传统媒体上难以获取的丰富信息。如果从中准确地分析和挖掘信息已经成为一个具有挑战性的研究问题。

概率主题模型被广泛地用于文档数据的潜在语义结构分析。经典主题模型假设文档由一个主题分布生成，其中每个主题是一个词典上的概率分布。经典主题模型已经成功地应用于新闻、科技文献和博客等文档集合。但是，当它们应用到 tweets、短消息以及论坛评论等短文本数据上时，结果却不够好。主要原因在于短文本和普通长文本相比缺乏词共现信息。为了解决这一问题，本文提出了两个方法，分别叫做词网络主题模型（WNTM）和伪文档主题模型（PTM）。

WNTM 通过从词-词空间学习主题缓解短文本的数据稀疏问题。促使我们这么做的一个原因是，当文档都很短时，词-文档空间是极其稀疏的，然而词-词空间仍然是稠密的。同时，从词-词空间学习主题更能够保证主题语义的一致性。因此，我们从短文本构造得到的词共现网络学习主题，而不是直接从短文本集合学习主题。我们给出了从短文本构建词共现网络的方法，并给出了如何通过 Gibbs 采样算法从网络中学习主题。

PTM 通过将短文本聚合到潜在伪文档中来创造额外的跨短文本词共现信息。因此，PTM 的关键是引入的伪文档，它可以隐式地将短文本聚合起来对抗数据稀疏问题。通过将建模短文本生成过程转化为建模较长的伪文档生成过程，PTM 能够更准确有效地进行参数估计。为了消除伪文档和主题之间非必要的关联关系，我们基于 PTM 提出了带有稀疏先验的伪文档主题模型（SPTM）。

现有短文本主题模型的一个潜在缺陷是它们均采用词袋模型的假设。该假设虽然可以提高模型计算的效果，但是忽视了文档中词序信息。而词序是准确学习短文本主题的重要信息。我们提出了伪文档 N-gram 主题模型（PTNG）。PTNG 和 PTM 一样，也通过短文本自聚合解决短文本数据稀疏问题。此外，PTNG 可以自动从短文本中学习 collocations。Collocation 的学习可以为 PTNG 引入词序信息，这进一步保证了短文本主题学习的准确性。

总的来说，我们提出了一些针对短文本主题建模的新方法。在解决短文本内容稀疏问题时，我们避免了额外信息的利用和依赖。这保证了新方法广泛的应用价值。此外，

我们的工作是首次尝试在短文本主题建模中打破词袋模型的假设，引入词序信息。和普通文本的主题模型研究相比，短文本主题建模的研究仍然处于初级阶段。我们希望本文的工作可以推进该领域的进步和发展。

关键词：短文本，主题模型，潜在狄利克雷分配，伪文档，词共现网络，N-gram

Abstract

With the spur of various kinds of Web applications, especially the explosively growth of online social media such as Twitter and Facebook, short texts have become the prevalent format for information of Internet. For instance, around 319 million active users in Twitter can generate nearly 500 million tweets everyday. This huge volume of short texts contain sophisticated information that can hardly be found in traditional information sources. Hence accurately discovering knowledge behind these short texts has been recognized as a challenging and promising research problem.

Probabilistic topic models have been widely used to automatically extract thematic information from large archive of documents. Standard topic models assume a document is generated from a mixture of topics, where a topic is a probabilistic distribution of words. Standard probabilistic topic models have achieved great success in modeling text collections like news articles, research papers and blogs. However, the results are mixed when they are applied directly to short texts such as tweets, instant messages and forum messages. The reason is mainly due to the lack of word co-occurrence information in each short text as compared to regular-sized documents. To handle the lacking of word co-occurrence information issue, we propose two methods namely Word Network Topic Model (WNTM) and Pseudo-document-based Topic Model (PTM).

WNTM alleviates the data sparsity of short texts by learning topics from word-word space. The main idea of WNTM comes from the following observation. When texts are short, word-by-document space is extremely sparse, while word–word space is still rather dense. Since the topic quality can be guaranteed in the dense word–word space, we conjecture to learn topic components from word co-occurrence network rather than document collection is more reliable. We show how to convert given short texts into a word co-occurrence network, and how to apply Gibbs sampling to reveal topics from the constructed network instead of original short texts.

PTM aggregates short texts into latent pseudo documents, which creates additional cross short text word co-occurrence information. The key of PTM is the introduction of pseudo documents for implicit aggregation of short texts against data sparsity. In this way, the modeling of topic distributions of tremendous short texts is transformed into the topic modeling of much less pseudo documents, which could be beneficial for parameter estimation in terms of both ac-

curacy and efficiency. To further eliminate undesired correlations between pseudo documents and latent topics, we also propose a Sparsity-enhanced PTM (SPTM) by applying Spike and Slab prior to topic distributions of pseudo documents.

One potential limitation of existing short text topic models is that they all follow the *bag-of-words* assumption. This assumption brings computational efficiency, but also ignores word order that might severely hurt the accuracy of topic modeling on short texts. We propose a novel topic model titled Pseudo-document-based Topical N-Gram model (PTNG). PTNG can leverage much less pseudo documents to self aggregate tremendous short texts, which helps to gain advantages in learning topic distributions on short texts. Besides, PTNG can also automatically determine unigram words and collocations based on context and assign topics to both individual words and collocations, which guarantees the accuracy of learned topics.

In summary, we propose a series of novel methods for short text topic modeling. By solving the data sparsity issue without relying on any auxiliary information, our methods can be applied to a wide range of applications. Besides, our study is the first attempt to make short-text topic modeling go beyonds the *bag-of-words* assumption. Anyway, compared to normal text topic modeling, the research of short text topic modeling is still at a preliminary stage. We hope this thesis can promote research in this field.

Key words: Short Texts, Topic Modeling, Latent Dirichlet Allocation, Pseudo Document, Word co-occurrence network, N-gram

目　录

第一章 引言 …… 1

1.1 研究背景 …… 1

1.2 研究现状 …… 3

1.3 本文工作 …… 5

1.3.1 研究目标与内容 …… 5

1.3.2 研究成果 …… 5

1.4 论文组织 …… 6

第二章 主题建模综述 …… 7

2.1 主题模型简介 …… 7

2.1.1 潜在语义分析 …… 7

2.1.2 概率主题模型 …… 7

2.1.3 非概率主题模型 …… 9

2.1.4 小结 …… 9

2.2 统计推断 …… 9

2.3 评价指标 …… 12

2.3.1 混淆度 …… 12

2.3.2 主题一致性 …… 13

2.3.3 间接评价 …… 14

2.4 主题模型的扩展 …… 14

2.5 结论 …… 16

第三章 词网络主题模型 …… 17

3.1 引言 …… 17

3.2 相关工作 …… 18

3.3 模型描述 …… 19

3.3.1 词共现网络 …… 19

3.3.2 词网络主题模型 …… 21

3.3.3 推断文档的主题 …… 22

3.4 复杂性分析与词网络调权 …… 22

3.4.1 复杂性分析 …… 22

3.4.2 词网络调权 …… 23
3.5 实验结果与分析 …… 24
3.5.1 评价主题质量 …… 24
3.5.2 词语义相似度计算 …… 26
3.5.3 文档分类 …… 28
3.6 小结 …… 32
第四章 伪文档主题模型 …… 35
4.1 引言 …… 35
4.2 模型描述 …… 36
4.2.1 基础模型 …… 36
4.2.2 模型比较 …… 37
4.2.3 稀疏模型 …… 39
4.3 统计推断 …… 40
4.4 模型扩展与讨论 …… 42
4.5 实验结果与分析 …… 43
4.5.1 实验设置 …… 43
4.5.2 短文本分类 …… 45
4.5.3 主题一致性 …… 48
4.5.4 参数敏感性 …… 48
4.5.5 扩展模型的验证 …… 50
4.6 小结 …… 50
第五章 伪文档 N-gram 主题模型 …… 51
5.1 引言 …… 51
5.2 N-gram 主题模型相关工作 …… 52
5.3 模型描述与统计推断 …… 52
5.3.1 模型描述 …… 52
5.3.2 模型讨论 …… 55
5.3.3 统计推断 …… 55
5.4 实验结果与分析 …… 57
5.4.1 实验设置 …… 57
5.4.2 实验结果 …… 59
5.4.3 短文本分类结果 …… 61
5.5 小结 …… 61

第六章 总结与展望 …… 63
6.1 本文工作总结 …… 63
6.2 未来工作展望 …… 64
参考文献 …… 65
攻读博士学位期间取得的学术成果 …… 77
致谢 …… 79
作者简介 …… 81

插 图 目 录

1 14 年至 16 年 Twitter 的月活跃用户增长情况 2
2 PLSA 和 LDA 的盘子表示法 8
3 生成过程和统计推断对比图，图片来源 [1] 10
4 定长滑窗以及词对自动加权示意图 20
5 WNTM 模型框架示意图 21
6 微博数据上词语义相关度计算任务上的等级相关系数结果 28
7 Wikipedia 数据上词语义相关度计算任务上的等级相关系数结果 29
8 WNTM 和基准方法在长短文本上的分类结果. 31
9 新闻正文分类的混淆矩阵 32
10 PTM 盘子表示法 36
11 PTM 和 PAM 的模型对比示意图 38
12 SPTM 盘子表示法 39
13 SPTM 盘子表示法 42
14 SATM 过拟合的图示 46
15 调整训练集大小的分类结果 47
16 主题一致性随伪文档数量的变化 48
17 PTNG 的盘子表示法 53
18 News 以及 DBLP 上 UCI 主题一致性结果 60

表 格 目 录

1 微博数据集上的主题一致性结果 25
2 Wikipedia 数据集上的主题一致性结果 26
3 新闻标题和正文每类文档数统计 30
4 非均衡语料的分类结果 33
5 数据集的统计指标 43
6 五折交叉验证的短文本分类结果 45
7 New 和 DBLP 上的 UCI 主题一致性结果 48
8 PTM、SPTM 和 EPTM 的主题一致性结果 49
9 PTM、SPTM 和 EPTM 的分类结果 49
10 PTNG 模型的数学符号表 54
11 数据集的统计指标 57
12 五折交叉分类实验结果 60

第一章　引言

从 Web1.0 到 Web2.0, 互联网完成了由静态页面展示到用户与网站动态交互信息的转变。其中社交媒体的兴起与蓬勃发展，更是极大地降低了用户发布消息的门槛。具体来说，社交媒体更注重人与人之间的交流，它的用户发布的内容较为随意也比较简短，例如 Twitter[1] 直接限制每条微博大小不能超过 140 字节。因此，越来越多的人能够轻松地参与到互联网信息的产生过程中去。这些用户产生的内容（User Generated Content，UGC）通常具有简短、非正式、主观性强和海量等特点，为传统文本挖掘算法的应用带来困难，同时也为文本挖掘研究带来了新的机遇和挑战。

1.1　研究背景

万维网（World Wide Web，WWW）的发明使得信息能够以 HTML 页面表示，并以超链接的方式开放给用户。只需连接到互联网，普通用户就可以通过浏览器方便地获取自己感兴趣的信息。在 Web1.0 时代，用户通常仅能够从互联网获取信息，也就是说信息的流通是单向地从网站到用户。到了 Web2.0 时代，用户可以参与到网站内容的建设中去，信息开始在用户和网站之间双向流通。比如，在线问答网站上的问题和回答都由用户编写和提交，维基百科上的词条也基本是由用户编辑。因此，互联网上用户产生的内容开始逐渐增多，其中包括了经过认真编辑的长文本信息，如维基百科和博客等，还包括了很多短文本信息如问答文本和用户评论等。

随着社交媒体的兴起和蓬勃发展，越来越多的用户成为社交媒体内容的贡献者。他们发布和分享信息的热忱被极大地激发出来，与此同时，互联网上用户产生的内容开始爆炸式增长。图 1展示了 14 年至 16 年 Twitter 的月活跃用户增长情况，可以看到 16 年年底，Twitter 的月活跃用户数量已经达到了 3.19 亿。除了活跃用户数目庞大之外，Twitter 上用户每日产生的微博数据也极为海量。据 2016 年的一份统计报告称，Twitter 上的用户每天合计产生约 5 亿条左右的微博。正如前文所说，社交媒体对用户发布的内容并无太多限制，甚至某些社交媒体约束用户不能提交过长的文本。这很大程度上降低了用户发布消息的门槛，导致了用户产生内容，特别是短文本数量的飞速增长。

互联网上用户产生的海量文本蕴含着丰富的信息。比如，社交媒体上的文本记录了人们的所见所闻乃至所感，电子商务网站的用户评论反映了他们对公司或者产品的某些

[1] http://twitter.com

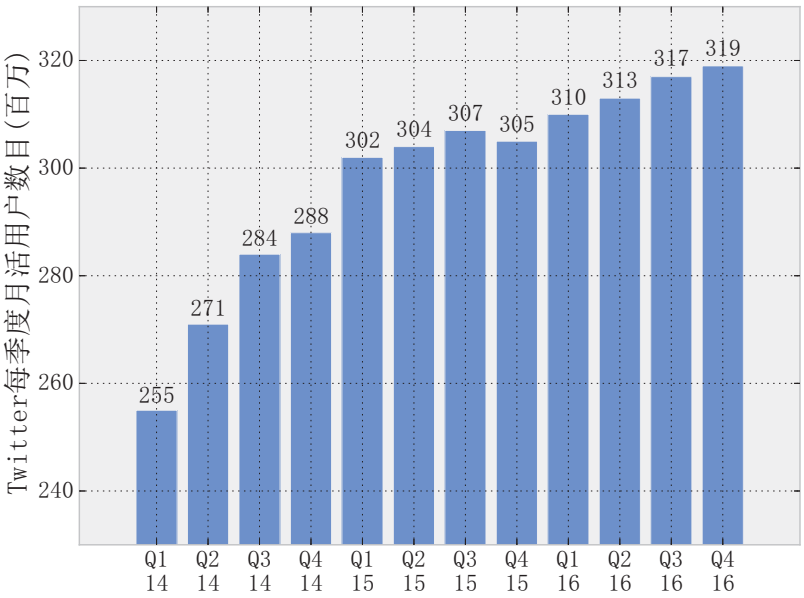


图 1　14 年至 16 年 Twitter 的月活跃用户增长情况

方面的观点，问答网站上的文本蕴含了各类经验知识。这些信息的分析与挖掘对许多应用都具有重要价值。下面我们给出一些例子：

- 用户观点分析 - 各类论坛、电子商务网站和社交媒体为互联网用户提供了发表和分享自己意见与看法的平台。通过对用户提交的评论、微博等短文本进行观点挖掘，一方面可以对用户个人的偏好进行分析去指导个性化推荐服务，另一方面可以整合群体的观点去预测选举的结果，收集每个产品的用户反馈等。
- 用户行为分析 - 用户除了分享自己的观点之外，还会在社交媒体上分享自己做了什么。这为研究人员提供了分析用户行为的第一手资料。例如，通过分析一些用户的饮食和运动习惯，研究人员可以分析他们的健康状况。
- 突发事件检测 - 当突发事件发生在自身周围时，一些社交媒体用户会在第一时间上传事件相关的消息。有研究表明，许多突发事件在被传统新闻媒体报道之前已经在社交媒体上传播了一段时间。这意味着我们可以通过分析用户提交的信息流来更早地发现突发事件。

虽然短文本蕴含着大量有价值的信息，但是分析和挖掘这些信息却绝非易事。首先，这些短文本极其简短。根据我们的实验，经过分词等预处理过程后，百度知道[2]的问题平均长度（即词的个数）为 4.2，Twitter 上微博的平均长度为 8.5。这意味着短文本和诸如新闻、论文和博客等长文本相比，有着特征极其稀疏，上下文缺失严重等特点。此外，短文本大多由普通用户产生，内容和格式均缺乏规范和约束。因此，社交媒体上的短文

[2]https://zhidao.baidu.com

本经常包含错字、新词和无意义的垃圾信息等。传统文本挖掘方法假设文本特征丰富，内容经过仔细编辑。这些假设在处理短文本时均不再一定成立，导致传统方法在短文本上通常不够有效。

1.2 研究现状

早期短文本语义处理和分析的一个重要研究方向是查询的理解和处理。用户提交给信息检索系统的查询通常包含 2-4 个词 [2]，属于典型的短文本。用词来表示查询，不利于查询和文档之间的匹配度计算，给基于向量空间模型 [3] 或者统计语言模型 [4] 的信息检索技术的应用带来不少的困难。如果一篇文档中不包含查询语句中的任何词，那么这篇文档和查询的匹配度就为零。查询特征稀疏和上下文缺失的特点必然导致搜索到的结果中缺失了许多与查询语义相关的文档。为了解决这个问题，在计算查询和文档的匹配度之前，研究者们利用查询文本中的 n-gram 或者 Wikipedia 等外部语料对查询的表示做扩展 [5, 6]，也有研究者提出先将查询提交给 Google[3]，然后利用返回结果中的摘要扩展原始查询 [7]。可以看出，信息检索领域解决查询的特征稀疏问题的一种方式是利用外部信息来增强短文本表示。

随着互联网技术的发展，特别是社交媒体的崛起，短文本逐渐成为互联网内容的重要组成部分。短文本的类型和相关应用变得多种多样，对它的语义分析受到越来越多研究者的关注。一些研究者在短文本聚类之前，首先利用外部信息（Wikipedia、Wordnet 或者其他开放网页）改善短文本的文档表示 [8–10]，获得了更好的聚类结果。Yan 等人 [11] 提出了基于跨文档词共现的词加权算法，提高了非负矩阵分解在短文本上聚类效果。关于短文本分类，Phan 等人 [12, 13] 提出首先用外部语料训练主题模型，然后用训练好的模型推断每条短文本包含的主题，最后用短文本包含的主题扩展它的词袋表示再训练分类器，最终提高了短文本分类效果。Yu 等人 [14] 用短文本的短语特征来扩展词特征，开发了一个针对短文本的 SVM 分类器 LibShortText。

近年来，基于概率主题模型的短文本语义分析受到了研究者们的重视。概率主题模型是一类概率图模型。和传统的空间向量模型和统计语言模型不同，概率主题模型在语义空间上表示文档，而不是词空间。这一方面实现了文档表示的降维，另一方面挖掘出的主题是可解释的，能够帮助人们了高效地了解文档集的主要内容。然而，经典概率主题模型 Latent Dirichlet Allocation (LDA) [15] 直接应用到短文本上面临文档级词共现稀疏的问题。Tang 等人 [16] 甚至从理论上阐述了 LDA 的推导算法直接应用于短文本时学习到的主题质量欠佳。目前已经有不少的工作提出了多种方法来解决短文本主题建模时遇

[3]https://www.google.com.hk

到的特征稀疏的问题。Hong 等人 [17] 首先将微博按照用户或者关键词聚合成伪文档，然后在伪文档上训练 LDA，发现两种聚合方式均比直接在微博上训练 LDA 效果好，同时发现按照用户聚合微博得到的效果最好。Mehrotra 等人 [18] 尝试了更多种聚合短文本的方式，比如按照时间戳、Hashtag 和突发词等。他们发现按照 Hashtag 聚合的效果是最好的。短文本在聚合的过程中，词语的一些跨文档共现信息被创造出来，这一定程度上弥补了短文本文档级词共现信息稀缺的不足。但是，用户和 Hashtag 等额外信息并不一定容易获取，这为上述方法的广泛应用带来困难。Quan 等人 [19] 提出一种自聚合主题模型，不利用额外信息也可以对短文本实现自聚合 (self-aggregate)。除了利用额外信息对短文本聚合之外，还有一些研究人员直接将额外信息考虑到文本的生成过程中去。比如，Tang 等人 [20] 将微博的用户和 Hashtag 等上下文信息整合到主题模型的生成过程中，提出了多上下文主题模型。Jin 等人 [21] 解析微博文本中短链接指向网页的正文，然后利用这些长文本的语义信息去帮助短文本的主题挖掘。最近，通用短文本主题模型的研发得到了越来越多的关注。除了上文提到的自聚合主题模型 [19] 之外，一些工作借鉴一元混合模型（Mixture of Unigram 或 MU） [22] 的思想，假设一篇短文本的所有词属于同一个主题。这降低了主题模型的复杂度（参数数量），使得模型可以适用于短文本。Yan 等人 [23] 则改进了 MU 模型，提出了双词主题模型。他们打破了短文本的文档边界，直接建模共现词对，认为每个词对里的词由同一个主题生成。Lin 等人 [24] 将 LDA 模型狄利克雷 (Dirichlet) 先验替换为一种贝叶斯稀疏先验。总的来说，改善短文本处理的方法大致可以分为两大类：

1、 利用额外信息 - 额外信息包括 Wikipedia、WordNet 等的外部信息，微博的用户、Hashtag、时间戳以及短链接等上下文信息。外部信息虽然可以改善短文本语义分析和处理，但是过分依赖外部信息会破坏短文本原始的语义。利用上下文信息面临的问题是并不是所有类型的短文本都有上下文信息，这限制了这类方法的广泛应用。

2、 利用内部信息 - 打破短文本的文档边界，如短文本聚合可以增加词共现信息或者直接建模文档集合的词共现信息，这有助于提升短文本主题模型的效果。另外，考虑词之间的顺序，如引入短语特征可以扩充短文本表示。利用内部信息的优势在于没有引入异构数据，这完全避免了噪音信息，而缺点在于没有一个通用的准则作为指导。

虽然目前有学者提出几种针对短文本的通用主题模型，但是这些模型有着各自的局限。例如，一元混合模型和双词主题模型过于简单，对复杂短文本数据缺少拟合能力。带有稀疏先验的主题模型在实际训练中效果欠佳。Quan 等人的自聚合主题模型时间复

杂度很高，并且模型参数数量随着数据一起增长，很容易过拟合。另外，现有短文本主题模型仍然采用词袋模型的假设，忽视了词序对短文本语义分析的影响。虽然有直接对 LDA 扩展得到的非词袋主题模型，但是它们在建模短文本时面临更加严重的特征稀疏问题。

1.3 本文工作

1.3.1 研究目标与内容

本文的研究目标是根据短文本的特点设计合适的概率主题模型。概率主题模型自提出一直发展到现在，经过了十多年的时间。它的应用几乎涵盖了文本挖掘和信息处理的方方面面。但是，目前大部分主题模型的研究均是针对长文本，如科技文献、新闻、博客和百科等。尽管有一些分析短文本的工作，但是很多直接应用传统的主题模型，忽视了短文本自身的特点。短文本上限制传统主题模型发挥作用的特点包括特征稀疏、主观性强等。近两年，针对短文本的主题模型研究受到越来越多的关注，但是现有的模型有着自身的缺陷。为此，本文以分析和挖掘海量用户生成的短文本为背景，研究和设计针对短文本的主题模型，以提高短文本语义分析的水平。具体而言，本文的研究内容如下：

1、 内容稀疏 - 经典概率主题模型假设每篇文档对应一个主题分布，这在分析新闻、科技文献和网页等长文本时问题不大。但是，在处理短文本时，特征的稀疏导致估计短文本的主题分析很困难。简而言之，模型相对于数据而言过于复杂。因此，本文的一个研究内容就是设计新的主题模型，使之适用于内容稀疏的短文本。

2、 忽略词序 - 经典主题模型在建模文档时使用词袋模型，即假设词和词之间的生成过程是相互独立的。这种简化保证了训练过程的效率，却忽视了词序对主题建模的作用。词序的考虑有助于消除语义建模中的歧义，例如，Wallach 等人 [25] 指出两个语义不同的句子可以包含完全相同的词。短文本内容稀疏导致了词序对语义建模的作用增大。因为短文本很可能只包含一个句子，忽略词序可能导致对拥有相同词袋表示但是语义不同的短文本学习到相同的主题。

1.3.2 研究成果

通过对上述内容的研究，取得了下列成果：

1、 提出了词网络主题模型

短文本数据内容过于稀疏，导致传统主题模型效果欠佳。这主要是由于主题模型依赖文档内的词共现信息学习主题。针对这一点，我们提出一种新的主题模型，不依赖文档内的词共现信息，而是直接建模文档集合级别的词共现。这有效地避免

了短文本文档内特征稀疏的问题，提高了短文本主题建模的效果。此外，我们还发现在词空间进行主题分析可以缓解内容分布不均衡问题。

2、 提出了伪文档主题模型

词网络主题模型虽然可以一定程度上避免原始短文本面临的特质稀疏问题，但是它在词空间挖掘主题，失去了建模原始短文本主题分布的能力。因此，我们提出了伪文档主题模型。传统主题模型相对于短文本数据过于复杂，而现有的一元混合模型和双词主题模型又过于简单。伪文档主题模型的复杂度介于二者之间。因此，它更加适用于短文本。此外，伪文档主题模型通过短文本自聚合增加了词共现信息，这也有利于主题的学习。

3、 提出了伪文档 N-gram 主题模型

传统主题模型忽视了文本中词的顺序信息。而基于传统模型考虑词序信息的扩展模型并不能直接应用到短文本，因为它们在短文本上面临更严重的特征稀疏问题。因此，需要研究针对短文本的考虑词序的主题模型。现有短文本主题模型要么难以引入词序，要么自身建模短文本的效果就不够好。因此，我们提出一种伪文档 N-gram 主题模型，在伪文档主题模型的基础上引入词序信息。

1.4 论文组织

本文共分为六章，内容组织如下：

第一章介绍了研究背景和现状、本文的研究目标、内容以及取得的研究成果。

第二章从主题建模简介、统计推断、评价指标以及主题模型的扩展等四个方面对主题建模做了简单的介绍。

第三章讲述了词网络主题模型，包括模型的框架、参数估计、文档主题分布的推断以及真实数据集上的若干实验结果与分析。

第四章讲述了伪文档主题模型，包括模型的定义、参数估计、模型扩展的讨论以及真实数据集上的若干实验结果与分析。

第五章讲述了伪文档 N-gram 主题模型，包括模型的定义、参数估计以及真实数据集上的若干实验结果和分析。

第六章对全部的研究做了总结，并对未来的研究工作做了展望。

第二章　主题建模综述

本章首先介绍了几类经典的主题模型，然后给出了概率主题模型的一种常用统计推断算法，接着介绍了评价主题模型的方式。最后，介绍了主题模型在文本挖掘中的一些应用。

2.1　主题模型简介

文本表示是利用机器处理文本的首要问题。早期，文本表示通常采用向量空间模型 [3] 和统计语言模型 [4]。二者虽然理论上并不相同，但是它们均直接用词来表示文本。这种方式形式简单、计算容易，并且实际效果很好。然而随着文本分析研究的深入，人们开始意识到直接用词来表示文本存在诸多问题。首先，词空间的文本表示是高维稀疏的，不利于很多机器学习算法的训练。其次，这种表示无法建模同义词和一词多义的现象。于是，需要一种更高级的文本表示方式，既能刻画文本的语义信息，又是低维的。

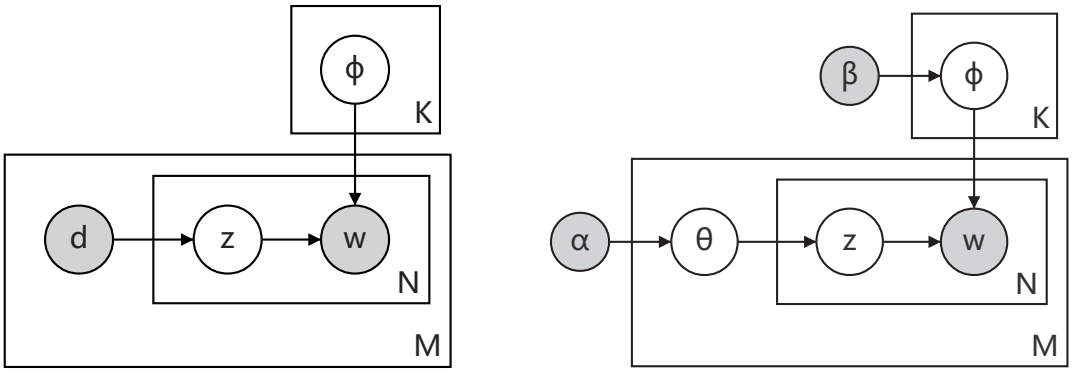
2.1.1　潜在语义分析

潜在语义分析 (Latent Semantic Analysis, LSA) [26] 给出了一种数据驱动的学习文档语义结构的方法，为文档语义表示学习的研究带来了一次飞跃。它假设文本数据中存在一个潜在的语义结构，通过对词 - 文档矩阵进行奇异值分解 (Singular Value Decomposition, SVD) 可以恢复潜在语义结构。LSA 利用 SVD 学习潜在语义暗含的假设是经常共现的词应该总是在同一个文档中出现。因此，LSA 实际上相当于对文档做了相关词扩充，一定程度上可以解决同义词的问题。

LSA 的提出为自动文本语义分析做出了开拓性的贡献，其利用词共现关系去学习潜在语义的思想启发了众多后续研究。然而，LSA 虽然可以解决同义词问题，但是无法解决一词多义。此外，SVD 的使用也限制了其应用到大规模的数据集中。

2.1.2　概率主题模型

Hofmann 在 1999 年提出了概率化潜在语义分析模型 (Probabilistic Latent Semantic Analysis，PLSA) [27]，从概率的角度重新定义了 LSA。PLSA 用生成模型描述了文档的生成过程，它将词和文档都看作随机变量，而且引入了一个隐变量 z 来表示词所属的潜在语义，称为主题（Topic）。PLSA 假设文档集合中有 K 个主题，给定一个文档 d，其中每个词 w 的生成过程如下：



(a) PLSA　　(b) LDA

图 2　PLSA 和 LDA 的盘子表示法

1. 从概率分布 $P(z|d)$ 中抽取一个主题 z
2. 从概率分布 $P(w|z)$ 中抽取一个词 w

此处的 $P(z|d)$ 和 $P(w|z)$ 均被定义为多项分布。模型的参数 $P(z|d), P(w|z)$ 可以通过最大似然的方式用期望最大化（Expectation-Maximization，EM）[28] 算法求解。

PLSA 用统计中的混合分解 (mixture decomposition) 代替了 LSA 中的 SVD。相比 LSA，PLSA 有着坚实的统计学基础。模型的生成过程和参数也有通俗易懂的概率解释。此外，PLSA 在实际使用时效果非常好并且时间复杂度更低，因此在工业界应用较多。

不过，PLSA 自身仍存在问题：

- PLSA 并没有描述文档的生成过程，导致它难以描述新文档的生成过程；
- 由 $P(z|d)$ 可知，PLSA 的参数数目随着文档数目的增加而增加，导致它更容易过拟合。

针对 PLSA 的不足，Blei 等人提出了 LDA（Latent Dirichlet Allocation [15]。LDA 为 $P(z|d)$ 和 $P(w|z)$ 引入了 Dirichlet 先验分布。Dirichlet 先验分布的引入一方面使得 LDA 能够描述新文档到主题的生成过程，另一方面减轻了 PLSA 容易过拟合的问题。在 LDA 中，一篇文档 d 有一个主题的多项分布 θ_d；而一个主题 k 是词的一个多项分布 ϕ_k。同时，θ_d 和 ϕ_k 分别由一个 Dirichlet 分布产生。具体的，文档集合的生成过程如下：

1. 为每个主题 $k \in [1, K]$ 采样一个词分布 $\phi_k \sim \mathrm{Dir}(\beta)$
2. 为每个文档 $d \in [1, D]$
    a. 采样一个主题分布 $\theta_d \sim \mathrm{Dir}(\alpha)$
        i. 采样一个主题 $z \sim \theta_d$
        ii. 采样一个词 $w \sim \phi_z$

图 2(a)给出了 PLSA 概率图模型表示（盘子表示法）。和图 2(b)中 LDA 的概率图表示相比，不难发现二者的差别在于 LDA 引入了 Dirichlet(α) 和 Dirichlet(β) 两个先验分

布。之所以选择 Dirichlet 分布作为先验，最主要的原因就是它和多项分布是共轭分布，可以方便计算后验概率进而简化参数的估计。除此之外，引入 Dirichlet 先验还有两个重要作用：1）缓解了 PLSA 的过拟合问题；2）可以描述一个新文档的生成过程。特别的，如果将 α,β 设置为 0，那么 LDA 等价与 PLSA。因此，PLSA 可以看作 LDA 的一种特殊情况。

2.1.3 非概率主题模型

非概率模型主要利用线性代数工具建模主题。目前主流方法基本采用非负矩阵分解 (NMF) [29] 而非 LSA 中的 SVD。NMF 将原始的词-文档矩阵分解为两个矩阵，并约束这两个矩阵为低秩非负矩阵。分解得到的两个子矩阵有着明确含义，分别代表词-主题矩阵和主题-文档矩阵。Gaussier 和 Goutte [30] 研究发现当使用 KL-divergence 作为损失函数时，NMF 和 PLSA 是等价的。因此 NMF 在实际应用中可以取得和 PLSA 类似的效果。

基于基本的 NMF，后续有很多扩展。Li 等人 [31] 总结了 NMF 的多种变型并提出了 Tri-factorization，将原始的词-文档矩阵分解为三个子矩阵，然后应用到文档聚类任务中。GraphNMF [32] 通过引入文档之间的相似度来约束矩阵分解。Zhu 等人 [33] 引入稀疏约束控制模型有效参数的数量。Wang 等人 [34] 提出了 GroupNMF 以协同学习不同类别下的主题。此外，也有部分工作为 LDA 引入约束条件以改善它的主题建模效果，如 Regularized LSI [35] 和 sparse LSA [36]。

除去以上两大类主题模型之外，还有一类基于神经网络的主题模型 [37–39]。它们采用无向图建模文档、词和主题之间的关系，有着表达方法更加灵活，但是训练代价高的特点。近年来，深度学习逐渐成为一个炙手可热的研究方向，已经有研究者开始尝试将深度学习方法引入文本语义分析与处理领域 [40, 41]，并取得了进展。

2.1.4 小结

概率模型和非概率模型各有优势。概率模型有坚实的统计学基础，模型模块化程度高，易于扩展。非概率模型少了概率约束，更为灵活，例如加入正则化约束，自定义词的权重等。正因为此，非概率模型比概率模型的原则性差些，效果调优较为麻烦。故而，本文将关注于概率主题模型，而不再讨论非概率模型。

2.2 统计推断

概率主题模型均是利用统计推断方法去学习模型的参数。统计推断可以看作是生成过程的逆过程，即基于观测到的文档集合去推断模型的参数。如图 3所示，左边的图展

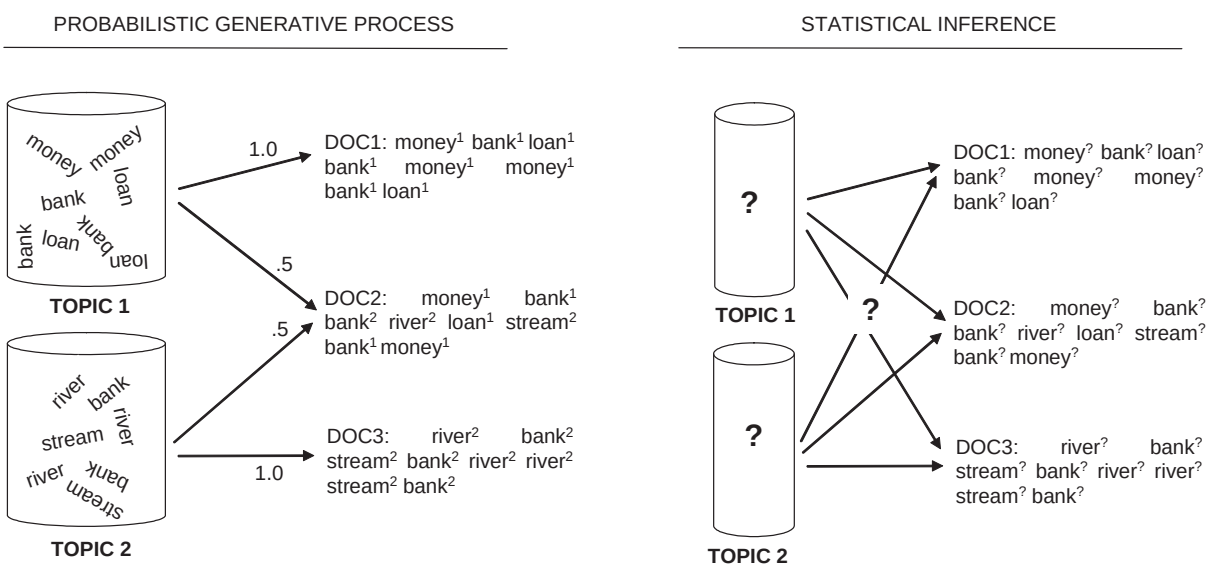


图 3　生成过程和统计推断对比图，图片来源 [1]

示了三篇文档的生成过程，在生成过程中，假设 ϕ 和 θ 是已知的。具体地，DOC1 中的词全部从主题 1 中抽取，DOC2 中的词有 0.5 的概率从主题 1 抽取，有 0.5 的概率从主题 2 抽取，DOC3 的词全部从主题 2 抽取。右边的图则展示了统计推断的过程，实际上 ϕ 和 θ 均是未知，每次词所属的主题 z 也是未知的，唯一观测到的就是文档中的词。统计推断就是要根据已观测到的文档来推断每个文档的主题分布，主题中词的分布，以及每个词的主题来源。

主题模型常用的统计推断方法包括 Gibbs 采样 [42]、变分推断 [15]、EP(Expectation Propagation) 算法 [43] 和最大化后验估计 [44] 等。其中最为流行的是 Gibbs 采样和变分推断。实际上，这两种方法和最大化后验概率方法的主要差别在于先验的平滑程度 [45]。根据 [45] 的结论，Gibbs 采样通常比其他两种算法的求解精度更高，消耗内存也更少。相对于变分推断而言，Gibbs 采样的求解过程也更简单。所以在本文当中，我们主要介绍 Gibbs 采样推断方法，其他方法可以参考相关文献。

Gibbs 采样 [46] 是蒙特卡洛方法 (Markov chain Monte Carlo，MCMC) 的一种特例，适用于高维隐变量模型的参数估计。其基本思想是按马尔可夫链的方式交替地对待估计的随机变量进行后验采样，其中每次采样基于其他随机变量的赋值。接下来以 LDA 为例，介绍用 Gibbs 采样算法来做统计推断的过程。在 LDA 中，需要估计的随机变量包括 K 个主题到词的分布 $\Phi = \{\phi_k\}_{k=1}^{K}$，$D$ 个文档到主题的分布 $\Theta = \{\theta_d\}_{d=1}^{D}$ 和每个词 w 的主题赋值 z。不过，利用 Collapsed Gibbs 采样技术 [42]，ϕ 和 θ 在具体求解过程中可以被积掉，并不需要显式的采样。因此，我们只需要对 z 进行采样。

假设文档集合包含 D 个文档,其可以表示为一个长度为 N 的词序列 $\mathbf{w} = \{w_1, w_2, \cdots, w_N\}$。

其中词 w_i 所属的文档记为 d_i。Gibbs 采样需要对每个词对应的主题 z_i 进行采样。为了进行采样，关键是计算隐变量的条件概率分布 $P(z_i|z_{\neg i}, \mathbf{w})$ 的计算。根据贝叶斯公式：

$$P(z_i|z_{\neg i}, \mathbf{w}) = \frac{p(\mathbf{z}, \mathbf{w})}{p(\mathbf{z}_{\neg \mathbf{i}}, \mathbf{w})} \propto \frac{P(\mathbf{w}|\mathbf{z})P(\mathbf{z})}{P(\mathbf{w}_{\neg \mathbf{i}}|\mathbf{z}_{\neg \mathbf{i}})P(\mathbf{z}_{\neg \mathbf{i}})}. \tag{2.1}$$

其中：

$$\begin{aligned} P(\mathbf{w}|\mathbf{z}) &= \int P(\mathbf{w}|\mathbf{z}, \Phi)P(\Phi)d\Phi \\ &= \int \left(\prod_{n=1}^{N} P(w_n|z_n, \phi_{z_n})\right) P(\Phi)d\Phi \\ &= \int \prod_{k=1}^{K} \left(\frac{\Gamma(\sum_{w=1}^{W}\beta_w)}{\prod_{w=1}^{W}\Gamma(\beta_w)} \prod_{w=1}^{W} \phi_{k,w}^{n_{w,k}+\beta-1} d\phi_k\right) \\ &= \left(\frac{\Gamma(\sum_{w=1}^{W}\beta_w)}{\prod_{w=1}^{W}\Gamma(\beta_w)}\right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{W}\Gamma(n_{w,k}+\beta_w)}{\Gamma(n_{.,k}+\sum_{w=1}^{W}\beta_w)}, \end{aligned} \tag{2.2}$$

这里 $\Gamma(.)$ 是标准的 Gamma 函数，$n_{w,k}$ 表示词 w 从主题 k 中抽取的次数，$n_{.,k} = \sum_{i=1}^{W} n_{w_i,k}$。$P(\mathbf{z})$ 的计算如下：

$$\begin{aligned} P(\mathbf{z}) &= \int P(\mathbf{z}|\Theta)P(\Theta)d\Theta \\ &= \int \left(\prod_{n=1}^{N} P(z_n|\Theta)\right) P(\Theta)d\Theta \\ &= \int \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{k,d}^{n_{k,d}+\alpha-1} d\Theta \\ &= \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(n_k+\alpha_k)}{\Gamma(D+\sum_{k=1}^{K}\alpha_k)}. \end{aligned} \tag{2.3}$$

同样的，可以计算出 $P(\mathbf{w}_{\neg i}|\mathbf{z}_{\neg i})$ 和 $P(\mathbf{z}_{\neg i})$：

$$P(\mathbf{w}_{\neg i}|\mathbf{z}_{\neg i}) = \left(\frac{\Gamma(\sum_{w=1}^{W}\beta_w)}{\Gamma(\beta_w)^W}\right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{W}\Gamma(n_{w,k|\neg i}+\beta_w)}{\Gamma(n_{.,k|\neg i}+\sum_{w=1}^{W}\beta_w)}, \tag{2.4}$$

$$P(\mathbf{z}_{\neg i}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(n_{k|\neg i}+\alpha_k)}{\Gamma(D+\sum_{k=1}^{K}\alpha_k)}, \tag{2.5}$$

将2.2 - 2.5带入到2.1，并利用 Gamma 函数的性质 $\Gamma(x+1) = x\Gamma(x)$，消掉公共项后得到 Gibbs 采样所需要的条件分布：

$$P(z_i = k|z_{\neg i}, \mathbf{w}) = (n_{k|\neg i} + \alpha_k) \frac{n_{w,k|\neg i} + \beta_{w_i}}{\sum_{w=1}^{W} n_{.,k|\neg i} + \beta_w}. \tag{2.6}$$

现在给出 Gibbs 采样算法的具体步骤。首先，我们对所有词的主题赋值进行随机初始化；然后对每个词计算其主题条件概率分布，即2.6，并据此采样一个主题。依次迭代，

直到收敛。我们通过收集采样得到的样本，就可以估计出模型的参数:

$$\phi_{k,w} = \frac{n_{w,k} + \beta_w}{n_{.,k} + \sum_{w=1}^{W} \beta_w}, \tag{2.7}$$

$$\theta_{d,k} = \frac{n_k + \alpha}{D + \sum_{k=1}^{K} \alpha}. \tag{2.8}$$

更多的细节和完整的 Gibbs 采样算法请参考文献 [47]。

以上介绍的是求解 LDA 的最基本的 Collapsed Gibbs 采样算法。当主题数目 K 设置的比较小，例如 $K = 100$ 时，Collapsed Gibbs 采样算法的性能还是令人满意的。但是，当 K 很大时，通常就需要一些时间复杂度更优的方法。SparseLDA [48] 利用主题到词的分布的稀疏性改进了采样算法，提高了时间效率。AliasLDA [49] 则提出，当 K 很大时，文档到主题的分布是非常稀疏的，利用这个特点同样显著提高了采样算法的效率。除去直接优化采样算法的研究外，还有一些研究工作利用分布式计算求解 LDA，使得 LDA 能够被应用到大规模的文档集上 [50–52]。

2.3　评价指标

如何评价一个主题模型的好坏一直是一个开放性问题，目前仍没有一个统一的标准。主要因为：1）主题模型是一种无监督方法，数据中本身并没有主题标注信息；2）主题模型通常作为一种文本表示学习或者特征学习，而不像文本分类、信息检索等应用有一个明确的目标。因此，主题的好坏评价标准对于不同的应用而言也可能不一样。比如，对于文本分类而言，要求主题之间的区分度较高；而对于信息检索而言，可能要求主题结果更全面一些，以保证检索结果的召回率。下面介绍目前主题模型相关研究中比较常见的一些评价方式，并讨论各自的优点与不足。

2.3.1　混淆度

自从 PLSA 和 LDA 提出以来，一种典型的评价方式就是在训练模型之前先保留一部分数据，这部分数据不参与训练，然后用训练后的模型去拟合保留数据。最常用的指标有混淆度 (Perplexity) 和似然函数 (Likelihood) 等。由于这些指标大同小异，我们只介绍最常见的一种——混淆度。混淆度是统计语言模型中常用的一个指标 [53]，常被用来评估模型的泛化能力。令 $\mathbb{D}$ 表示保留的测试集文档集合，混乱度的定义如下：

$$\text{perplexity}(\mathbb{D}|\Phi,\Theta) = \exp\{\frac{\sum_{d\in\mathbb{D}} \log(P(\mathbf{w_d}|\Phi,\Theta))}{\sum_{d\in\mathbb{D}} N_d}\},$$

其中

$$\log(P(\mathbf{w_d}|\Phi,\Theta)) = \sum_{i=1}^{N_d} n_{d,w_i} \log(\sum_{k=1}^{K} \phi_{k,w_i}\theta_{d,k}),$$

其中，n_{d,w_i} 表示词 w_i 在文档 d 内出现的次数。混淆度越低，说明学习到的参数对测试集拟合的越好，证明模型的泛化能力更好。

基于测试集拟合度的评价方式的好处是不依赖于外部数据和人工标注信息，计算简单。然而，这种评价方式是有争议的。Chang 等人发现对测试集拟合度好的模型学习到的主题反而可能与人对主题质量的评价相左 [54]。最近 David Blei 也承认这种评价方式是和人们对主题模型的预期结果是脱节的 [55]。

2.3.2 主题一致性

近年来，不少研究者认为主题模型的出发点就是要学习到有语义清晰的主题。因此，应该直接评价主题的语义一致性，即主题中高概率的词之间的语义相关性越强，则该主题的可读性越好。

一种最简单的评价主题一致性的方式就是把每个主题中概率最大的几个词列出来，然后由人去对其相关性进行标注 [56]。但是这种方式主观性太强，很少使用。Chang 等人提出了一种词入侵和主题入侵的方式来分别评价主题和文档主题分布的质量 [54]。词入侵是在每个主题的概率较大的几个词中间随机插入一个其他主题中概率较大的词；主题入侵是在每个文档对应的主题分布中取前 3 个概率最大的主题，然后随机插入一个概率较小的主题。然后让标注者去找出侵入词或主题。如果侵入词或主题很容易被区分，说明结果较好。

但是，人工标注一方面成本较高，另一方面容易受到标注者的主观影响。因此，很多研究人员仍然在积极地探索数据驱动的自动评价主题一致性的方法。Newman 等人 [57] 在 2010 年提出用外部数据（如 WordNet，Wikipedia，Google 搜索结果等）来自动化地评价主题一致性。其主要思想是从外部数据中来计算主题中概率最大的几个词之间的语义相关性。其实验结果表明，利用大规模知识库或者网页数据，可以取得和人工评价相关性很高的结果。2011 年，Mimno 等人 [58] 提出了另外一种自动化主题一致性评价方法–coherence score。Coherence score 的特点是无需借助外部数据，只利用训练主题模型的文档集合内部的词共现关系去评价主题中概率最大的前几个词之间的语义相关性。一个主题中排在前面的词在文档集合中共现的文档频率越高越好。其实验结果表明 coherence score 评价结果与人工标注有较强的相关性。

相比前面两种方式，主题一致性评价更接近于主题建模方法的初衷，即学习到有人可以理解的语义清晰的主题，是值得提倡的一种评价方式。

2.3.3　间接评价

在实践中，主题模型常被用作降维的工具，辅助文档的分类、聚类和检索。即将文档从词空间降到主题空间，然后用降维后的语义特征辅助特定的应用。此时，应用所特有的评价指标会被用来评价主题建模的效果。这类评价称为间接评价，常见的有以下几种：

1、 文本分类

相关工作有 [15, 33, 59] 等。这种评价方式主要是将 θ_d 作为文档 d 的表示，然后用常用分类器，如朴素贝叶斯、SVM 等对其分类。分类效果越好，说明主题模型学习到的主题区分度越高，从分类角度讲主题建模的效果也越好。这种评价方式适合文档有类别标签的数据。

2、 文本聚类

相关工作有 [54, 59, 60] 等。这种评价方式认为同属于一个类的文档，降维后的相似度应该较大；相反，原先不属于同一个类的两个文档，降维后相似度应该较小。具体有两种做法：1）把每个主题看做是一个类，然后把每个文档 d 分配到它最可能采样的那个类，即 $P(z|d)$ 最大的那个主题；2）用 θ_d 作为文档 d 的表示，然后用 KNN 等常用聚类算法对其聚类。同样的，这种方式适合文档有类别标签的数据。

3、 文本检索

相关工作有 [27, 34] 等。基本思想是将查询和文档映射到语义空间，然后用语义表示来计算文档之间的相似度。检索结果越好，说明主题模型学习的结果对文档的表示越合理。

间接评价的优势就是与应用挂钩，能直接体现主题模型的实用价值。但由于主题模型本身并非针对某个应用的，这种评价方式相对片面。

2.4　主题模型的扩展

PLSA 和 LDA 是两个最基本的概率主题模型，后续有大量的工作是基于它们的扩展。如 LapPLSA [60] 通过引入基于文档之间的相似度构造正则化因子来约束主题的学习，以保证相似度高的文档其学习到的主题分布一致。考虑到 LDA 的扩展则非常多，这里我们只简单介绍几种有代表性的工作。Blei 等人在 LDA 的基础上提出了层次化主题模型 [61]，该模型能自动学习主题之间的层次结构。PLSA 和 LDA 都需要事先指定主题的数目，Teh 等人提出的 HDP 模型能用非参方法去自动学习主题的数目 [62]。Wang 等人结合主题模型和协同过滤算法做科技文献的推荐 [63]。部分研究工作将主题模型扩展到多个语言的数据集上 [64, 65]。部分研究工作将 LDA 与有监督学习结合起来，提出了一些有

监督的 LDA 模型 [66–69]。部分研究工作将元信息引入到 LDA 中来学习主题和这些元信息的关系。如 author-topic 模型 [70] 同时建模了作者信息，TOT 模型 [71] 建模了时间信息。部分研究工作将语法信息引入到 LDA 的学习中，同时考虑词的语义和语法信息 [72, 73]。部分研究工作将文档的链接关系考虑到了主题模型中 [74–76]。

主题模型的研究长期以来都是围绕长文本进行的，或者说，研究人员默认每篇文档有足够多的词可以供主题模型学习语义。但是，随着 Web 2.0 以及社交媒体的发展，如今的互联网上已经随处可见短文本。那么经典的主题模型如 PLSA [27] 和 LDA [15] 直接应用到短文本上时效果如何呢？Hong 等人 [17] 指出传统的主题模型 LDA 在 Twitter 数据上效果欠佳，他们还发现对微博进行聚合可以改善主题模型的效果。Tang 等人 [16] 从理论上证明了文档长度过短时，LDA 无法准确地学习主题。这些研究均说明了对短文本进行语义分析需要改进现有的主题模型。下面将具体介绍目前已有的一些工作针对主题模型在短文本上改进的工作。

某些类型的短文本虽然文本长度短，但是包含丰富的附加信息。比如微博中，除了文本内容，往往还包含用户 ID、hashtag、时间戳、地理位置和 URL 等。这些附加信息往往可以被用来改善主题模型。Hong 等人 [17] 发现利用用户 ID 对微博进行聚合后再训练 LDA 可以提高效果。Mehrotra 等人 [18] 尝试了多种聚合微博的方式，如按时间，按突发词，按 hashtag 等，其结论是按 hashtag 聚合的效果更好。除了直接聚合短文本，再利用经典主题模型训练数据的方法外。一些工作尝试直接将附加信息建模到概率主题模型中，开发新的模型。Tang 等人 [20] 提出了多上下文主题模型（multi-context topic model），认为有相似附加信息的微博在语义上也应该是接近的。Jin 等人 [21] 利用许多微博包含外部链接的特点，提出了主题知识迁移学习方法，对微博和微博中 URL 指向的网页正文进行联合主题建模。需要指出的是，不是所有的短文本都有附加信息。例如，新闻标题、查询语句和论坛评论等大部分都不包含附加信息。此外，只有部分微博都有 hashtag 或者 URL。因此，上述利用附加信息的方法并不算是解决短文本主题建模的通用办法。

除了利用附加信息的工作外，已经有一些旨在提出通用的短文本主题模型的研究。Yan 等人 [23] 指出混合语言模型（Mixture of Unigram，MU）比 LDA 更加适用于短文本，并且在 MU 的基础上提出了双词主题模型，减少了 MU 严格限制每篇短文本只有一个主题所引入的错误。Zhao 等人 [56] 也通过限定每条微博只能属于一个主题来改进主题模型。MU 这种一条短文本只能属于一个的主题的假设过于极端。Lin 等人 [24] 提出了对偶稀疏主题模型，将 LDA 的 Dirichlet 先验换成了稀疏先验，这样模型从理论上可以应对短文本只属于少数主题的情况。MU 可以看成是对偶稀疏主题模型的极端情况。Quan 等人 [19] 提出了自聚合主题模型，指出不依赖附加信息，将短文本聚合和主题建模放在

一个统一的模型里可以取得很好的效果。虽然双词主题模型以及对偶稀疏主题模型是通用的短文本主题模型，但是它们并没有增加词语的共现，因此并未实际解决短文本主题建模的困境。另外，虽然自聚合主题模型可以增加文档级的词语共现，但是模型本身存在容易过拟合的问题。

2.5　结论

主题模型在过去十年间一直是机器学习和数据挖掘领域中的一个热门研究话题。本文首先回顾了主题模型发展的历史，并着重介绍了概率主题模型及其统计推断方法。然后，本文总结了目前常用的几种主题模型评价方式，指出了各自的优点与不足。随后，本文例举了主题模型在文本挖掘的一些领域中的应用。从主题模型的广泛应用可以看到主题模型为文本的语义分析提供了一种自动化的工具，能有效提高计算机对大规模文本语料处理的水平，在实际应用当中具有广泛的应用价值和深远的意义。最后，本文也指出以往大部分的主题模型研究都是针对长文本。近几年，越来越多的研究人员开始关注如何有效的对短文本进行主题建模。然而，虽然已经有部分工作讨论了短文本主题建模，但是相关的研究才刚起步，还有待深入。

第三章　词网络主题模型

3.1　引言

随着社交媒体的发展，短文本已经成为互联网信息的重要表现形式。尽管大量短文本中包含着复杂且有用的信息，但是长度短、内容分布不均衡等特点限制了传统主题模型在短文本上的直接应用。传统主题模型如 LDA [15] 假设一篇文档 d 对应一个主题的多项分布 θ_d，每个主题 k 对应一个词的多项分布 ϕ_k。生成文档中一个词 w_i 时，LDA 首先通过 θ_d 采样一个主题 z_i，然后通过 ϕ_{z_i} 采样 w_i。采用统计推断算法，可以估计 LDA 的参数 θ_d 和 ϕ_k。一般而言，主题模型通过文档级词共现信息将语义相近的词归到同一个主题下。这导致主题模型对文档的长度以及主题在文档上分布的均衡性较为敏感。因为短文本包含很少的词，同时词频很低，主题模型学习面临词共现信息稀疏的问题。此外，LDA 之类的主题模型也很难发现仅出现在少数文档中的主题，因为它们倾向于学习在文档中分布比较广的主题。最近 Tang 等人 [16] 从理论上证明了，当文档数目不足时，LDA 无法准确学习主题。这意味着当一个主题分布在很少的文档上时，LDA 之类的模型将很难学习出这个主题。

目前有许多工作尝试解决 LDA 建模短文本时效果不好的问题。例如，Weng 等人 [77] 在训练 LDA 之前将有关联的短文本聚合成伪文档。Phan 等人 [12] 在通用大规模语料（如 Wikipedia）上训练主题模型，之后用训练好的模型推断短文本中每个词对应的主题。此外，还有一些针对短文本分析的基于 LDA 的扩展模型 [56, 78, 79]。上面提到的方法要么依赖于特殊数据，要么针对特定应用，这限制了它们的广泛应用。最近通用短文本主题模型的研究受到越来越多的关注。一个典型的例子是双词主题模型 [23]（Biterm Topic Model 或 BTM）。虽然 BTM 在短文本上取得了比 LDA 等传统模型更好的效果，但是它本质上是一种一元混合模型（Mixture of Unigram 或 MU）。这意味着 BTM 模型过于简单，可能无法拟合复杂的短文本数据。另一个短文本主题模型是对偶稀疏主题模型 [24]（Dual Sparse Topic Model 或 DSTM）。DSTM 将 LDA 的狄利克雷先验分布替换为一种贝叶斯稀疏先验，使得模型能够灵活地建模短文本之类的稀疏数据。然而，DSTM 的模型复杂度高于 LDA，这导致了 DSTM 在短文本上的实际效果比较有限。

当主题在文档上分布不均衡时,一些工作利用额外的先验知识指导主题的学习 [80, 81]，还有一些工作使用非均衡狄利克雷先验 [82]。值得注意的是，在实践中关于待处理数据的先验知识通常是未获取的或者不容易获取的。另外，虽然采用非均衡狄利克雷先验在

一定程度上可以缓解主题分布不均衡的问题，但是先验对模型的约束力有限。

为了同时解决短文本的内容稀疏和主题分布不均的问题，在本章中，我们提出了词网络主题模型（Word Network Topic Model 或 WNTM）。和 LDA 生成文档不同，WNTM 通过生成词共现网络来发现主题。提出 WNTM 主要基于如下几个考虑：

1、 当所有文档都很短时，词-文档空间是非常稀疏的，但是词-词空间还是稠密的。因此，我们认为从词共现网络中学习主题可以避免短文本的特征稀疏问题。实际上，处理短文本时，从词-词空间学习到的主题一致性是有理论保证的 [83]。

2、 尽管一个主题可能分布在很少的文档上，但是它所关联的词可能并不少。因此，从词共现网络学习主题有利于稀有主题的发现。

3、 研究一个适用于短文本，并且能够学习稀有主题的简单而又通用的方法是有实际需要的。

具体而言，WNTM 利用 Gibbs 采样算法从词共现网络中发现词团（即主题），并且学习一个词的邻接词表在词团上的分布。学习词而非文档的主题分布，有助于 WNTM 避免短文本的特征稀疏和主题分布不均的问题。虽然 WNTM 不再建模文档的主题，但是通过词的主题仍然可以间接得到文档的主题表示。词共现网络可以从任何类型的文本中构建，这保证了 WNTM 可以应用于多种多样的实际问题。

本章剩余部分组织如下：3.2节介绍了相关工作；3.3节介绍了 WNTM；3.4节分析了 WNTM 的复杂度并提出了词网络调权算法；3.5节给出了实验结果和分析；3.6节给出了本章小结。

3.2　相关工作

概率主题模型如 PLSA [27] 和 LDA [15] 被广泛地应用于文本分析领域。相比于 PLSA，LDA 是一个更完整的生成模型。因为 LDA 有先验分布，而 PLSA 没有。由于 PLSA 和 LDA 这类概率图模型的高可扩展性，在过去的十多年里，大量基于它们的扩展模型被提出。例如，考虑了主题随着时间演化的动态主题模型 [84]，考虑了社交关系的社交主题模型 [85] 以及利用作者的主题生成文档的作者主题模型 [70]。这些扩展工作中的大部分都是针对普通长文本，同时考虑一些额外的元信息。

近年来，针对短文本特征稀疏问题的研究越来越多。早期的工作主要通过短文本聚合或者引入外部信息来帮助主题建模。比如，Hong 等人 [17] 首先利用用户信息或者相同的关键词将微博聚成较长的伪文档，然后训练 LDA。他们发现和直接在短文本上训练 LDA 相比，在伪文档上训练得到的模型效果更好。Jin 等人 [21] 解析微博文本中短链接指向网页的正文，然后利用这些长文本的语义信息去帮助短文本的主题挖掘。然而，短

文本聚合仍缺乏原则性的指导，利用外部语料可能引入过多噪声。改进的主题模型也是应对短文本特征稀疏的一种方式。Zhao 等人 [56] 假设每个微博只对应一个主题，微博中的所有词都由这个主题生成。Yan 等人 [23] 则改进了一元混合模型，提出了双词主题模型。Lin 等人 [24] 分别为文档-主题分布和主题-词分布引入了贝叶斯稀疏先验，以改进诸如短文本等稀疏文本上的主题建模效果。这些特定的短文本主题模型要么假设过强，要么模型过于复杂，在实际应用中的效果提升有限。

对于不均衡文本的主题建模研究则主要依赖于先验知识的利用。Andrzejewski 等人 [81] 提出了狄利克雷森林先验（Dirichlet forest prior），该先验能够建模 must-links 和 cannot-links 两种知识。Must-links 是指两个词应该出现在同一个主题中，cannot-links 意味着两个词应该出现在不同的主题中。Chen 等人 [80] 利用通用的词汇知识（lexical knowledge）学习一致性（coherence）更好的主题。值得注意的是，文档级别的知识也是可以利用的。Ramage 等人 [67] 和 Rubin 等人 [86] 设计的生成过程除了生成文档中的词，还同时生成了文档的标签。Blei 和 McAuliffe [66] 提出了有监督主题模型，也可以生成类别信息。这些利用了额外知识的方法均可以一定程度上缓解稀疏主题难以被学习到的问题。无论是词语级别还是文档级别的知识，在特定应用场景下，这些额外信息的获取通常是不容易的。除了这类方法，不均衡先验比均衡先验更有助于 LDA 学习到稀疏主题 [82]，但是先验的约束对最终学习到的模型影响有限。

和上述方法不同，我们尝试找到一种简单但是通用的方法，同时解决短文本上的特征稀疏以及主题分布不均的问题。

3.3 模型描述

传统主题模型利用文档中丰富的词共现信息去发现主题，然而短文本恰恰缺少文档级词共现信息。另外，主题模型的目标是最大化生成文档集合的概率，因此稀有的主题很容易被忽略掉。因此，传统主题模型直接应用于短文本或者不均衡文本时效果并不理想。为了解决上述问题，我们提出了词网络主题模型（Word Network Topic Model，WNTM）。WNTM 通过 Gibbs 采样算法，从词共现网络中学习词团（即主题）。此外，WNTM 学习词而非文档的主题分布。下面，我们首先介绍如何构建词共现网络，然后介绍 WNTM 的生成过程和采样算法，最后说明如何得到一篇文档的主题表示。

3.3.1 词共现网络

在词共现网络中，结点是文档集合中的词，两个结点之间存在边意味着两个词在相同的上下文中共现过。这里的上下文指的是文档或者构建词共现网络的定长滑动窗口。

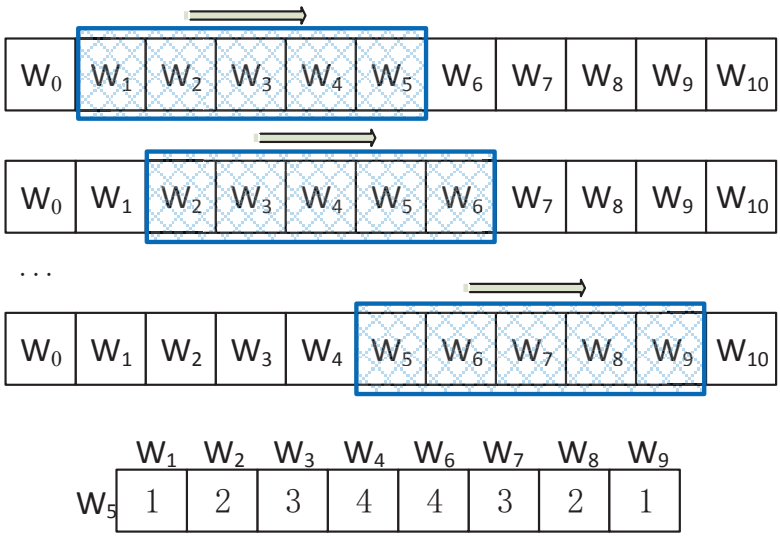


图 4　定长滑窗以及词对自动加权示意图

这里为了限制词共现网络的大小，我们采用定长滑窗做为上下文。对于长文本，过长的滑窗，会导致生成的词共现网络规模规模很大，进而导致主题计算速度很慢。幸运的是，根据 Peirsman 等人的工作 [87]，长度为 10 个词的滑窗足够获取词的语义相关性。对于短文本，以定长滑窗或者直接文档为上下文均可。结点的度（degree）定义为它的相邻边的权重之和，而结点的活跃度（activity）定义为它的相邻边的平均权重。

为了将文档集转换成词共现网络，我们首先去除文档集中的低频词以及停用词，然后用长度为 10 的滑窗逐个词地扫一遍文档。随着窗口的滑动，窗口内任意两个词均被认为共现了一次。我们将词对共现的次数累加起来，得到的和被定义为词对所对应边的权重。随着滑窗的移动，文档中一个词对的共现次数可能被多次统计。这个性质我们称之为词对的自动加权。具体而言，一个词对离的越近，它们的共现次数被重复统计的越多。如图 4所示，词 W_5 和词 W_4（或 W_6）共现了 4 次，而词 W_5 和词 W_1（或 W_9）仅共现了 1 次。这种加权方式是合理的，因为词和距离近的词语意更相关。这有助于 WNTM 学习到语义更一致的主题。

在基于文档的主题模型中，主题通常是指经常在相同文档中共现的一组词。从这一点来看，主题和词共现网络中词团是类似的。因为词共现网络中连接紧密的词对一定经常出现在同样的滑窗中，进而它们有很大概率出现在相同的文档中。因此，我们可以将词共现网络中的词团看成传统主题模型的主题。其实，处理短文本时，从词共现网络中学习主题理论上是可以保证主题一致性的 [83]。此外，稀有主题相关的词也是有可能在词共现网络形成一个紧致的词团，这有利于基于词共现网络的主题模型发现稀有主题。基于以上考虑，我们提出词网络主题模型 WNTM。为了保证 WNTM 的简单和可扩展性，

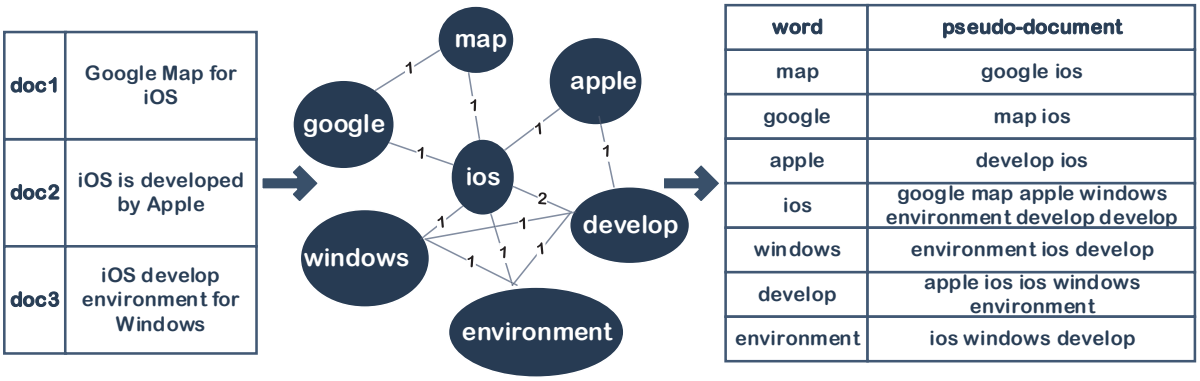


图 5　WNTM 模型框架示意图

我们采用 Henderson 和 Eliassi-Rad 的策略 [88] 去学习词团，即主题。

3.3.2　词网络主题模型

LDA 的 Gibbs 采样算法可以直接应用于网络的社团发现 [88]。不过，在应用 Gibbs 采样算法前，需要先将网络表示为结点的邻接表。将每个结点的邻接表看作伪文档，我们可以直接在这些伪文档上应用 LDA 的 Gibbs 采样算法，得到的主题就是一个社团。采用这个思路，我们也首先将词共现网络转化为伪文档。我们假设词共现网络的边是加权无向的。如图 5所示，网络中的每个词都对应一个伪文档，而伪文档的内容就是词对应的邻接词列表。因为词共现网络是加权的，所以邻接词可能在伪文档里出现多次。

尽管 WNTM 使用和 LDA 一样的 Gibbs 采样算法，但是它们生成过程的目标是不一样的。LDA 学习如何用主题生成文档，而 WNTM 学习如何用词团生成词的邻接词表。更具体地说，WNTM 通过假设词的邻接表由一个概率图模型生成来学习的词、词团、词的邻接词之间的关系。WNTM 首先假设词共现网络中有一组数量固定的词团，每个词团 z 是在所有词上的多项分布 ϕ_z，而该分布本身从一个狄利克雷 $Dir(\beta)$ 采样得到。全部伪文档的生成过程可以如下描述：

1、 为每一个词团 z 采样一个在所有词上的多项分布 $\phi_z \sim \mathrm{Dir}(\beta)$

2、 为词 w_i 的词邻接表 L_i 采样一个在所有词团上的多项分布 $\theta_i \sim \mathrm{Dir}(\alpha)$

3、 为邻接表 L_i 中的每个词 w_j:

（1） 采样一个词团 $z_j \sim \theta_i$

（2） 采样一个词 $w_j \sim \phi_{z_j}$

在 WNTM 中，θ 表示词的邻接表中的词属于某个词团的概率，ϕ 表示词属于某个词团的概率。给定一个文档集合，WNTM 首先将它转换为词共现网络，然后将网络转换成伪文档，最后利用 LDA 的 Gibbs 采样算法估计所有的 ϕ 和 θ。和 LDA 不同，WNTM

并不学习文档的主题分布，而是学习词的上下文的主题分布。因为伪文档，即词的邻接表，包含了词的所有上下文信息。

3.3.3 推断文档的主题

如上节所述，WNTM 并不建模文档的生成过程，因此它无法直接得到文档的主题分布。不过，由于 WNTM 建模词的邻接词表的生成，所以它可以得到词上下文的主题分布。这里我们将词 w_i 上下文 L_i 的主题分布 θ_i 当作词 w_i 自身的主题分布。有了所有词的主题分布，我们可以间接地计算文档的主题分布。

假设文档 d 包含 N_d 个词，根据链式法制：

$$P(z|d) = \sum_{i=1}^{N_d} P(z|w_i)P(w_i|d), \tag{3.1}$$

其中 $P(z|w_i)$ 就是 WNTM 学习到的 $\theta_{i,z}$。至于 $P(w_i|d)$ 可以通过经验估计得到：

$$P(w_i|d) = \frac{n_d^{w_i}}{N_d}, \tag{3.2}$$

其中 $n_d^{w_i}$ 是词 w_i 在文档 d 中的词频。因此，有了 θ 和公式 3.2就可以通过公式 3.1间接得到文档的主题分布。

3.4 复杂性分析与词网络调权

尽管 WNTM 和 LDA 在统计推断时使用同样的 Gibbs 采样算法，但是它们的时空复杂度并不一样。本节我们先比较 LDA 和 WNTM 的模型复杂度，然后提出一种词网络调权的算法降低 WNTM 模型的复杂度。

3.4.1 复杂性分析

LDA 的时间复杂度是 $O(N_d K_z L_d)$，其中 N_d 是文档数，K_z 是主题数，而 L_d 是平均文档词数。类似的，WNTM 的时间复杂度是 $O(N_p K_g L_p)$，其中 N_p 是伪文档数，即文档集的词典大学，K_g 是词团数，即主题数，而 L_p 是平均伪文档词数。假设滑动窗口长度为 c，因为文档集中最多有 $N_d L_d$ 个滑动窗口，并且每个滑动窗口可以产生 $\binom{c}{2}$ 条边，所以有如下式子成立：

$$N_p L_p \approx N_d L_d c(c-1). \tag{3.3}$$

Algorithm 1 词网络调权算法

Require: 原始词共现网络 $G = (V, E, W)$，其中 V 代表所有词的集合，E 代表所有边，W 代表边的权重.

Ensure: 调权后的词共现网络 $G' = (V, E, W')$。

1: 为 V 中的每个结点 n 计算它的度 $D(n)$ 和活跃度 $A(n)$
2: **for all** $e = (n_1, n_2) \in E$ **do**
3: 重新计算设置权重 $w_e = \left\lceil \frac{w_e}{A(n_i)} \right\rceil, \underset{i}{argmin}\{D(n_i),\ i = 1, 2\}$
4: **end for**

如果假设 K_z 等于 K_g，那么可以得到 WNTM 的时间复杂度是 LDA 的 $o(c^2)$ 倍。在实践中，因为短文本的平均文档长度 $\langle l \rangle$ 很接近 c，所以 WNTM 在短文本上的时间复杂度是可以接受的。但是，对于长文本而言，设置一个大一些的 c，将导致 WNTM 的训练非常缓慢。

LDA 的空间复杂度是 $O(N_d K_z + N_d L_d)$，而 WNTM 的空间复杂度是 $O(N_p K_g + N_p L_p)$。如果我们假设 $N_d = N_p$ 并且 $K_z = K_g$，那么 WNTM 的空间复杂度也将是 LDA 的 $o(c^2)$ 倍。为了将 WNTM 的时空复杂度降到 LDA 的线性复杂度，我们提出一种词网络调权方法。这种方法可以有效地提升 WNTM 的训练速度，并降低它的内存消耗。

3.4.2 词网络调权

上面的复杂性分析表明了直接在原始词共现网络生成的伪文档上应用 Gibbs 采样算法的时空复杂度是较高的。为了减少 WNTM 的复杂度，我们需要降低 $N_p L_p$ 的值。因为 N_p 的值是固定的，等于文档集的词典大小，所以 L_p 成为唯一可以调整的参数。如上文所述，L_p 等于伪文档的长度。从词网络的角度看，L_p 还等于结点的度。因此，如果可以降低网络的边权，就可以减少伪文档的平均长度，相应的，就可以降低 WNTM 模型的时空复杂度。在调整词共现网络权重的同时，还要保持不同词之间连接的紧密程度。

如算法 1所示，通过将边 e 的权重 w_e 除以度 $D(n)$ 相对小的端点 n 的活跃度 $A(n)$，可以得到边 e 新的权重。这么做，我们可以降低整个词共现网络的全部边权之和，进而降低 WNTM 的时空复杂度。因为结点 n 的度 $D(n) > c - 1$，所以词共现网络的平均度一定远大于 $c - 1$。因此，伪文档的平均长度应该小于 $\frac{L_p}{c-1}$，其中 L_p 是调权前词共现网络生成的伪文档。由公式 3.3可知，调权后的 WNTM 的时空复杂度是 LDA 的 $o(c)$ 倍。因此，上面的词共现网络调权算法可以将 WNTM 的复杂度降到 LDA 的线性复杂度。这保证了 WNTM 可以被有效地应用于短文本甚至长文本。

3.5　实验结果与分析

本节中，我们从三个方面对 WNTM 进行评价，包括主题的质量、词语义相似度计算和文档分类。每个方面的实验都分别采用了真实的短文本和长文本数据。对于短文本，我们采用 LDA [15]、双词主题模型 BTM [23] 和对偶稀疏主题模型 DSTM [24] 做为基准方法。对于长文本，我们对比 WNTM 和 LDA。和 LDA 在长文本上的对比，一方面可以说明 WNTM 可以应用于长文本，另一方面可以说明从词共现网络和文档集合学习主题各自的优劣。

本章的大部分实验均运行于一台配置了一块 2.40GHz 英特尔赛扬处理器和 12G 内存的 Windows 服务器上。由于 Wikipedia 数据集的数据量很大，相应的实验运行于一个由 13 台 Linux 机器组成的集群上。每台 Linux 机器有两块 2.27Ghz 英特尔赛扬处理器和 12Gb 内存。关于 LDA 和 WNTM 的实现，在短文本上我们使用一个 Java 的 LDA 开源实现 JGibbLDA[1]。在长文本上，我们使用一个基于 MPI 的开源工具 PLDA[2]。BTM 使用的是其作者开源的代码[3]。DSTM 由于没有开源实现，我们自己实现了它的代码。实验中，主题数目 K 统一设置为 100，模型 Gibbs 采样的迭代次数统一设置为 2000。JGibbsLDA 和 BTM 的 α 设置为 $50/K$，β 设置为 0.01。PLDA 的 α 设置为 0.1，β 设置为 0.01。DSTM 的参数设置和其论文上保持一致 [24]。除了新闻文档的分类实验，其他实验结果均是 10 次重复实验结果的均值。

3.5.1　评价主题质量

评价主题模型的常用方法是在测试文档集上计算模型的混淆度（perplexity）。由于 WNTM 并不建模文档的生成过程，所以混淆度并不适用于它。此外，最近的研究发现混淆度和主题的可读性并不是正相关的 [54]，具体而言，混淆度越低（越好）并不意味着主题的可读性越高。因此，这里我们采用名为主题一致性（Topic Coherence）的指标 [58]。

主题一致性

主题一致性通过计算主题高概率词之间的语义相关性评价主题的质量。主题一致性的值越大意味着主题的质量越好，即主题更容易被人理解。主题一致性的定义如下：

$$C(z;M^{(z)}) = \sum_{t=2}^{T}\sum_{l=1}^{t-1} \log \frac{D(m_t^{(z)}, m_l^{(z)}) + \epsilon}{D(m_l^{(z)})}, \tag{3.4}$$

[1] http://jgibblda.sourceforge.net/
[2] http://code.google.com/p/plda/
[3] http://code.google.com/p/btm/

表 1　微博数据集上的主题一致性结果

T	5	10	20
LDA	−36.6±1.8	−221.4±4.0	−1484.6±39.4
DSTM	−36.5±1.9	−199.6±7.7	−1260.5±29.7
BTM	−37.4±1.9	−207.5±8.2	−1235.9±30.7
WNTM	**−32.5±1.3**	**−181.6±5.5**	**−1056.6±22.8**

其中 $M^{(z)} = (m_1^{(z)}, ..., m_T^{(z)})$ 是主题 z 的 T 个概率最大的主题词，$D(m)$ 是包含词 m 文档个数，$D(m, m')$ 是同时包含词 m 和 m' 的文档个数。$\epsilon = 10^{-12}$ 的作用是防止分子出现为零的情况。上面的公式计算了单个主题的一致性，我们计算 K 个主题一致性的均值作为模型最终的主题一致性结果。

短文本上的主题一致性

为了验证 WNTM 能够在真实的短文本上学习到高质量的主题，我们从新浪微博采集了一天的微博数据。和 Twitter 类似，新浪微博也限制每条微博的长度不超过 140 个字节。由于微博文本内容不是很正规，我们对它进行了仔细的预处理。首先，我们采用 NLPIR[4]分词器对所有微博切词。然后，去除了停用词和文档频率小于 20 的低频词。接着，去除了 URLs、话题标签（Hashtag）、表情符以及非中文字符。最后我们仅保留词数大于等于 10 的微博。经过预处理，最终微博数据集包含 189223 个微博和 20942 个词。微博的平均词数为 17.2。

在微博数据集上，我们选择 LDA、BTM 和 DSTM 作为基准方法。表 1列出了主题一致性结果，其中 T 分别取 5，10 和 20。我们可以发现 WNTM 的主题一致性明显高于其他方法。进行 Mann-Whitney U 测试得到 p-value < 0.001，这说明了 WNTM 的性能提升具有统计显著性。和 LDA 相比，WNTM 学习到了更好的主题。这个结果说明了处理短文本时从词-词空间学习主题的优势。特别的，WNTM 的话题质量也优于 BTM 和 DSTM。这说明了 WNTM 和专门为短文本设计的其他模型相比也具有优势。虽然 BTM 和 WNTM 一样直接建模词共现，但是 BTM 由于模型过于简单导致其拟合数据的能力有限。此外，WNTM 能够通过滑窗为共现词对加权，这也是 WNTM 比 BTM 学习到更一致主题的部分原因。和 DSTM 的对比也验证了，稀疏先验的引入虽然使得模型的假设更贴合短文本稀疏的特点，但是同时也导致了模型更加复杂。这最终导致 DSTM 实际学到的主题质量不如直接从词-词空间学到的好。

[4]http://ictclas.nlpir.org/downloads

表 2　Wikipedia 数据集上的主题一致性结果

T	5	10	20
LDA	−13.9±0.3	−69.7±1.0	−327.4±3.3
WNTM	−13.8±0.2	−69.5±0.6	−329.7±2.4

长文本上的主题一致性

为了验证 WNTM 在实际的长文本数据上也能学到不错的主题, 我们在 Phan 等人 [12] 公开的 Wikipedia 数据上分别训练 LDA 和 WNTM。Wikipedia 数据集包含 71986 篇文档以及 60649 个词，它的平均文档长度为 423.5。表 2列出了主题一致性结果，和短文本上的实验一样，T 分别取 5，10 和 20。不难发现，当 $T = 5$ 和 10 时，WNTM 的主题质量好过 LDA。而 $T = 20$ 时，LDA 的主题质量稍好于 WNTM。WNTM 通过局部词共现信息学习主题，这有助于前 10 个高概率主题词之间的语义一致性。但是，当高概率主题词取的更多时，可能需要文档级别的词共现信息来保障主题词之间的一致性。不过，从结果来看 WNTM 和 LDA 学到的主题质量差别并不大。根据 Mann-Whitney U 测试的结果，WNTM 和 LDA 的主题一致性结果并无统计意义上的显著差异性。因此，对于长文本而言，从词-词空间学习到的主题质量和词-文档空间学到的差别不大。

3.5.2　词语义相似度计算

因为主题一致性是一种评价主题的内部指标，所以主题一致性高并不意味着模型在外部任务上也能有好的效果。例如，LDA 的主题一致性比 LSA 的好，但是 LSA 学习到的词或文档的语义表示在外部任务上比 LDA 的效果好 [89]。因此，我们进一步在词语义相似度计算和文档分类等外部任务上对比 WNTM 和基准方法的表现。这有助于展示模型在学习词或文档的语义表示时的有效性。下面首先介绍如果计算词对的语义相似度, 接着介绍词语义相似度计算任务，最后给出结果和分析。

对于 LDA、BTM 和 DSTM 而言，词 w 到主题的条件概率就可以作为词的语义表示

$$s_w = [p(z_1|w), p(z_2|w), ..., p(z_K|w)],$$

其中 K 代表主题数。当 Gibbs 采样算法收敛后，很容易通过经验估计计算得到 $p(z|w)$:

$$p(z|w) = \frac{n_{w|z}}{n_w}, \tag{3.5}$$

其中 $n_{w|z}$ 代表生成过程中词 w 被分配到主题 z 下多少次，n_w 代表词 w 在文档集中一共出现多少次。由于 WNTM 直接建模词上下文的生成，所以 θ_w 可以直接作为词 w 的语义表示：

$$s_w = \theta_w = [\theta_{w,1}, \theta_{w,2}, ..., \theta_{w,K}].$$

因为词的语义表示实际上是在主题上的分布，所以我们可以通过 Jensen-Shannon（JS）散度计算词对之间的距离。

$$\mathrm{JS}(s_i, s_j) = \frac{1}{2} D_{kl}(s_j \parallel m) + \frac{1}{2} D_{kl}(s_i \parallel m), \tag{3.6}$$

其中 s_i 和 s_j 分别是词 w_i 和词 w_j 的语义表示，$m = \frac{1}{2}(s_i + s_j)$，$D_{kl}(p\|q) = \sum_i p_i \ln \frac{p_i}{q_i}$ 是 KL 散度。如果我们将词语义表示看作空间中的点，那么夹角余弦也可以用来计算词对之间的距离：

$$\mathrm{Cosine}(s_i, s_j) = \frac{s_i \cdot s_j}{\|s_i\| \|s_j\|}. \tag{3.7}$$

词语义相关度计算任务常被用来评价语义空间。主题模型学到的主题就构成了一个词到主题的语义空间。如果可以准确地训练一个主题模型，那么我们可以认为语义相关的词对有相似的语义表示。在微博数据上，我们使用 Wang 等人 [90] 提出的评价中文词对的语义相关度计算任务。在 Wikipedia 数据上，我们采用两个评价英文词对的语义相关度计算任务 [91, 92]。在每个语义相关度计算任务中，均有若干词对的语义相关度经过了人工评分。分数越高意味着词对的语义越相关。Wang 等人给出了 240 个人工评分过的词对。Finkelstein 等人 [91] 给出的任务包含 353 个人工评分过的词对，包含了 Rubenstein 和 Goodenough [92] 给出 65 个词对。

我们首先根据公式 3.6或者公式 3.7计算词对的 JS 散度或者余弦相似度。然后计算所有词对的相似度和它的人工评分之间的 Spearman 等级相关系数（ranked correlation）。这里需要注意的是 JS 散度表示的是分布之间的距离，所以 JS 散度越大代表词对之间相似度越低。而夹角余弦正好相反，结果越大意味着词对相似度越高。等级相关系数结果越高说明模型学到的词语义空间越准确。图 6展示了在微博的短文本数据上的等级相关系数结果。从结果中可以发现，所有为短文本设计的主题模型在微博数据词语义相关度任务上均超过了 LDA，无论是根据 JS 散度还是余弦相似度。这说明了加稀疏先验或者直接建模词共现均可以提高主题模型学习词语义的准确性。而短文本主题建模方法中，WNTM 的表现是最好的（根据 Mann-Whitney U 测试，p-value < 0.001）。

图 7展示了在微博的短文本数据上的等级相关系数结果。我们惊讶地发现 WNTM 仍然可以超过 LDA。不过两者的差距不如在短文本上明显，这是因为长文本上 LDA 不

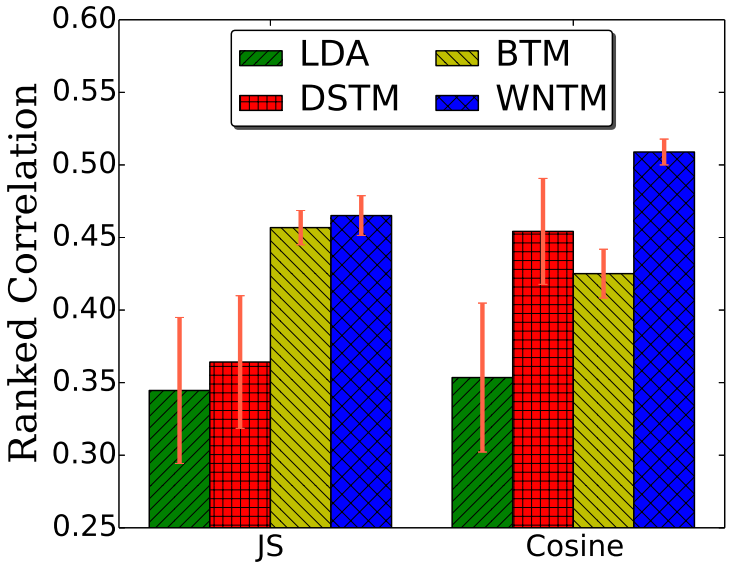


图 6　微博数据上词语义相关度计算任务上的等级相关系数结果

再面临数据稀疏的问题。WNTM 能够超过 LDA 可能和 WNTM 直接建模词的主题分布有关。尽管 WNTM 和 LDA 在主题一致性上结果非常接近，但是 WNTM 在词语义相关度计算任务上有明显优势。

3.5.3　文档分类

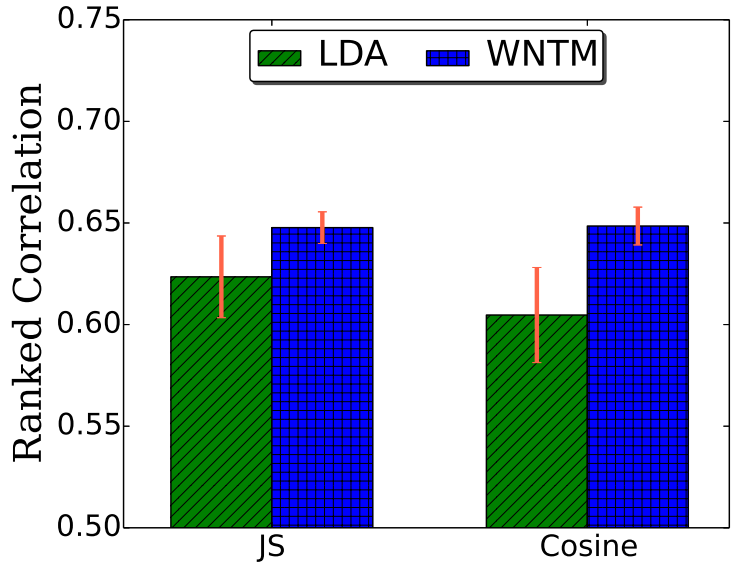
为了对比不同主题模型在学习长短文本语义表示上的能力，我们在搜狗新闻标题和新闻正文的语料上进行文档分类实验。为了验证 WNTM 能够更好地学习稀有主题，我们进一步对比了 WNTM 和 LDA 在非均衡语料上的分类结果。

新闻文档分类

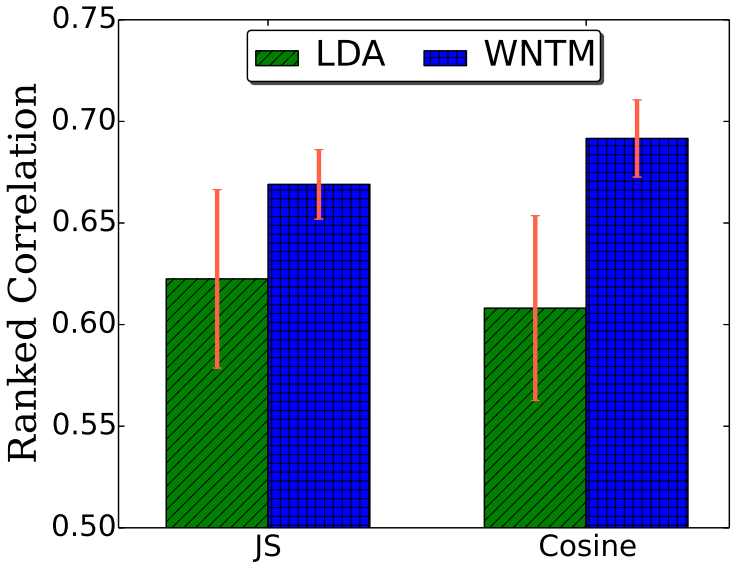
为了验证 WNTM 在学习长短文本语义表示上的能力，我们从 Sogou.com 下载了新闻语料[5]。经过预处理，我们最终得到了 508554 条新闻标题和一个大小为 59348 的词典。预处理后新闻标题的平均长度为 5.5。除了新闻标题，我们还获得了 118705 条新闻正文，其词典大小为 76114，正文平均长度为 175.9。每条新闻都有一个类别，新闻标题和新闻正文每类文档数的统计结果列在表 3中。

主题模型是一种降维方法，它可以将一篇论文表示到主题空间，即得到文档的主题分布。将每个主题在文档上的分布看作分类特征，我们可以根据文档的特征和类别训练分类器。具体地，我们分别在新闻标题和正文上训练主题模型，得到文档的主题分布

[5]http://www.sogou.com/labs/dl/ca.html



(a) Finkelstein 等人提供的任务上的结果



(b) Rubenstein 和 Goodenough 提供的任务上的结果

图 7　Wikipedia 数据上词语义相关度计算任务上的等级相关系数结果

后，采用十折交叉验证的方法训练和评价分类器。这里的分类器使用的是 LIBLINEAR[6]，一种线性核的 SVM 分类器。分类结果评价指标用的是宏平均的 precision、recall 和 f-measure。图 8(a)给出了新闻标题上的分类结果。图 8(b)则给出了新闻正文上的分类结果。

从新闻标题的分类结果中可以发现，DSTM 和 LDA 相比，它学习的主题表示可以

[6]http://www.csie.ntu.edu.tw/~cjlin/liblinear/

表 3　新闻标题和正文每类文档数统计

(a) 新闻标题每类文档数统计

类别	文档数	类别	文档数
Finance	31,414	Car	6,532
Sports	25,414	IT	2,321
Society	14,889	Military	1,733
Ent	11,208	House	1,410
Lady	8,128	Culture	983
Olympics	7,117	Health	962
Education	6,594		

(b) 新闻正文每类文档数统计

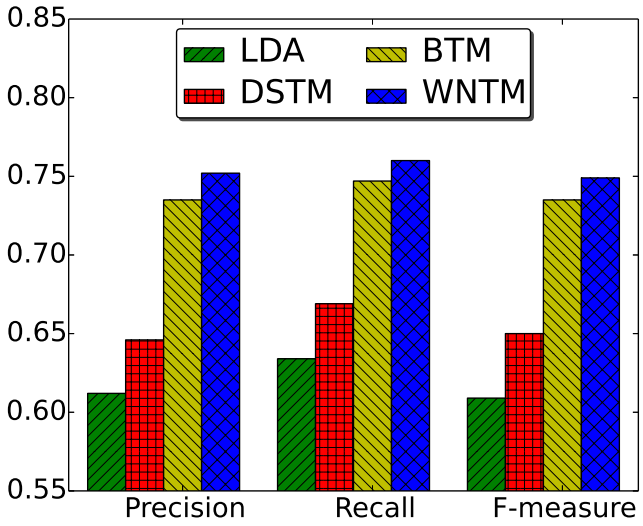
类别	文档数	类别	文档数
Finance	133,480	Car	18,675
Sports	115,946	IT	10,650
Society	70,743	Military	8,706
Ent	53,335	House	6,407
Olympics	34,767	Health	2,340
Lady	31,689	Culture	2,334
Education	19,482		

训练更准确的分类器，这意味着文档-主题分布的稀疏先验可以帮助主题模型学到更有区分度的文档主题分布。不过，DSTM 的结果比 BTM 和 WNTM 的差很多，这说明了在学习文档主题分布上，稀疏先验不如直接建模词共现有效。WNTM 的结果是最好的，超过了 DSTM 和 BTM（根据 Mann-Whitney U 测试，p-value < 0.001）。根据新闻正文的分类结果，LDA 的分类结果和 WNTM 的差距缩小了。和词语义相似度计算任务上的结果类似，尽管 WNTM 和 LDA 在主题一致性上结果非常接近，但是 WNTM 在学习更有区分度的文档主题分布上较 LDA 有明显优势。

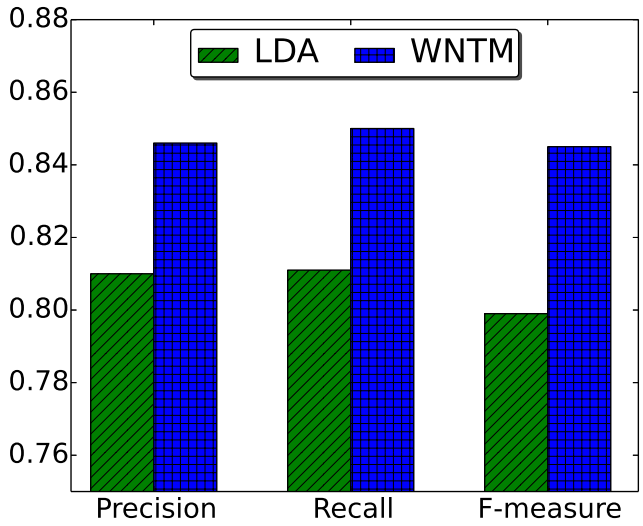
图 9展示了 LDA 和 WNTM 在新闻正文分类上的混淆矩阵。从图 9(a)中的结果可以发现，LDA 容易将类别为 “House” 的文档误分为 “Sports”、“Finance” 和 “Lady”。虽然 “House” 和 “Finance” 语义上较为相关，容易导致分类不准，但是直观而言 “House” 和 “Sports” 或者 “Lady” 并无明显的语义关系。结果 LDA 错分了 80% 以上的 “House” 下的文档。类似的情况还出现在了 “Culture”、“Health” 的分类结果中。回顾表 3可以发现，这三个错分比较多的类别都是稀有类别。这个结果间接说明了 LDA 缺乏对稀有主题建模的能力。从图 9(b)的结果可以看到，WNTM 对上述类别的分类结果比 LDA 好很多。

不均衡文档分类

Sogou 新闻分类实验中 WNTM 比 LDA 更准确地分类了稀有类别的文档。这一定程度上说明了 WNTM 比 LDA 更容易学习到稀有主题。为了进一步验证这一点，我们构造了一个可以调节某个类别稀有程度的语料。首先，我们创建均衡语料。均衡语料从上面介绍的新闻正文中抽取 5 个类别的新闻正文，各 1000 篇。然后，我们从类别 “Car” 中随机地去掉一些文档。通过控制删除的文档数量，可以调节 “Car” 类别文档的稀有程度，进而调节



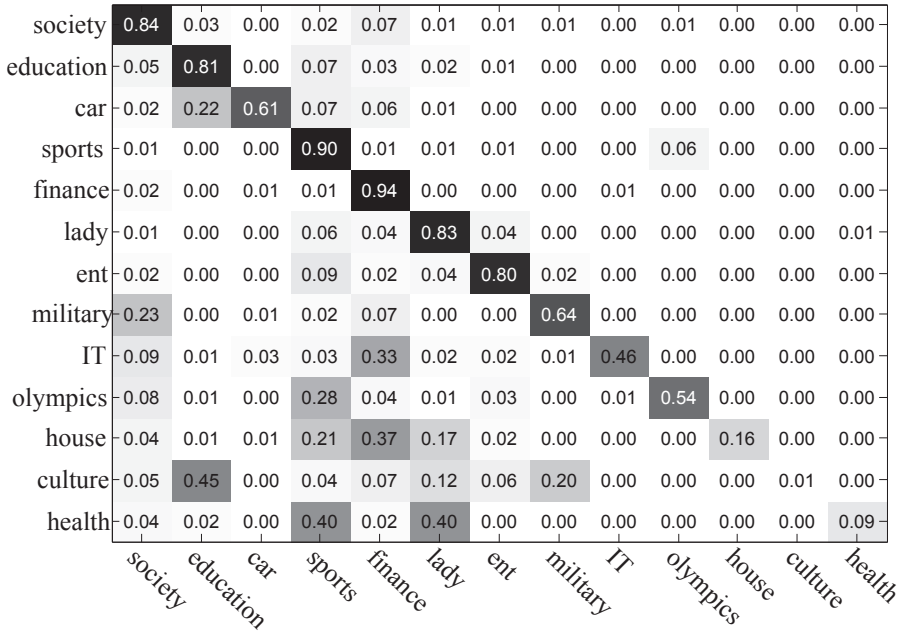
(a) 新闻标题上的分类结果



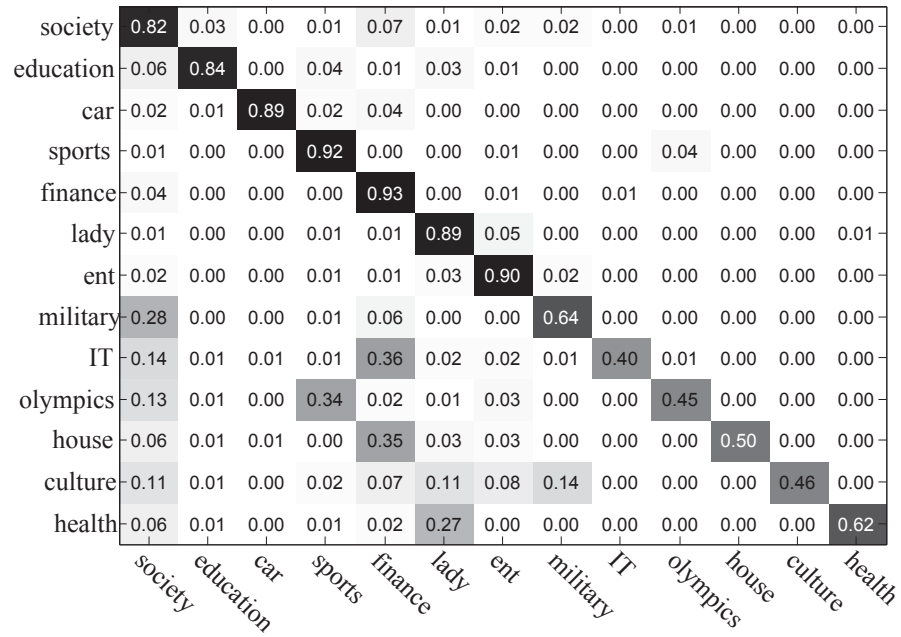
(b) 新闻正文上的分类结果

图 8　WNTM 和基准方法在长短文本上的分类结果.

语料的不均衡程度。具体地, 我们创建了 8 个不均衡语料。每个语料除了 “Car”, 其他类别均包含 1000 篇新闻正文。而 “Car” 的文档数目 $d_c = \{800, 600, 400, 200, 100, 80, 60, 40\}$。随着 d_c 的减少，语料的不均衡程度越来越高。为了消除删除文档随机性的影响，每个 d_c 对应的语料都重复生成了 10 份。表 4中的结果是 10 次结果的均值。从结果中可以发现, 随着 d_c 从 800 降到 200，LDA 的准确率先是稍稍上升，随后快速下降。相反的，WNTM 的准确率一直降的比较慢。特别的，当 d_c 等于 60 和 40 时，LDA 的准确率已经分别降到了 87% 和 76%，而 WNTM 的准确率分别是 88% 和 82%。这个结果说明了 WNTM 比 LDA 更加准确的学习到了稀有类别文档的主题表示。虽然当 d_c 在 500 到 200 之间时,



(a) LDA



(b) WNTM

图 9　新闻正文分类的混淆矩阵

LDA 的准确率高于 WNTM，但是它的召回率和 F 值均不如 WNTM。总之，通过构造非均衡语料的分类实验，我们证实了 WNTM 比 LDA 更适合建模稀有主题。

3.6　小结

本章中我们提出了一种简单通用的短文本主题方法，即词网络主题模型（WNTM）。通过生成词共现网络，WNTM 不仅仅可以从短文本中学习到一致性更强的主题，还能

表 4　非均衡语料的分类结果

d_c	Precision		Recall		F-measure	
	LDA	WNTM	LDA	WNTM	LDA	WNTM
800	0.948±0.009	**0.958±0.008**	0.881±0.012	**0.941±0.010**	0.913±0.007	**0.950±0.005**
600	0.953±0.007	0.957±0.007	0.864±0.011	**0.923±0.012**	0.906±0.007	**0.936±0.006**
400	0.956±0.011	0.954±0.006	0.839±0.026	**0.911±0.016**	0.893±0.015	**0.932±0.009**
200	**0.961±0.012**	0.932±0.011	0.792±0.021	**0.866±0.019**	0.868±0.014	**0.898±0.013**
100	**0.935±0.019**	0.909±0.022	0.727±0.043	**0.792±0.038**	0.818±0.034	0.846±0.025
80	0.903±0.069	0.913±0.023	0.718±0.087	0.773±0.042	0.798±0.077	0.836±0.029
60	0.874±0.039	0.884±0.034	0.619±0.199	0.700±0.100	0.702±0.186	0.789±0.051
40	0.760±0.100	0.820±0.065	0.491±0.192	0.553±0.119	0.600±0.200	0.658±0.104

一定程度上提升主题模型对不均衡文档中稀有主题的学习能力。通过在真实长短文本上的实验，我们发现 WNTM 不仅在主题一致性这样的内部指标上超过了基准方法，它还在词语义相关度计算和文档分类等外部任务上获得了最好的效果。这说明了 WNTM 可以学习到更好的词语义空间或者文档表示。特别的，在非均衡分类实验中，我们验证了 WNTM 比 LDA 更好地建模稀有类别文档，间接证明了 WNTM 对稀有主题的学习能力。

本章的研究成果已于 2016 年被数据挖掘期刊 Knowledge and Information System（KAIS）录用，论文题目为 “Word network topic model: a simple but general solution for short and imbalanced texts”。

第四章　伪文档主题模型

4.1　引言

随着短文本主题建模受到越来越多的关注，不断有新的短文本主题建模方法被提出。其中一些方法在建模短文本时，取消了文档边界的限制，直接建模文档集的共现词对[23]。上一章提到的 WNTM 就属于这类方法。另一些方法通过引入相关长文档、元数据等额外信息加强短文本主题模型[20, 21]。还有一些方法通过修改模型提升短文本主题建模效果，例如假设短文本只有一个主题 [56] 或者采用稀疏先验等方式 [24]。虽然上述方法一定程度上改善了短文本主题建模的效果，但是它们均没有增加额外的词共现信息，而词共现信息缺少恰恰是短文本主题模型学习效果差的主要原因。并且，上一章介绍的 WNTM 由于在词空间挖掘主题，失去了建模原始短文本主题分布的能力。因此我们需要研究一种能够增加词共现信息并能够建模原始短文本主题分布的主题模型。

一些研究者在训练 LDA 之前利用用户、时间或者话题标签等额外信息将短文本聚合成长的伪文档 [17, 18, 77]。这种在短文本上训练主题模型的方式虽然有着额外信息难以获取，方式过于启发式等缺陷，但是却是唯一一类增加了词共现信息的方法。当两个短文本合并后，原本不在同一篇短文本的词一同出现在了合并后的文档中。这意味着伪文档集合上的词共现信息多于原来短文本集合上的，即用来训练主题模型的词共现信息增加了。最近，Quan 等人 [19] 提出了自聚合主题模型（SATM）。尽管 SATM 不依赖外部信息，仅靠短文本自身的主题进行短文本自聚合，但是它的参数个数随着数据增加而增加。这导致它很容易过拟合。此外，SATM 的采样算法时间复杂度非常高。SATM 的缺陷限制了它在实践中的应用。

因此本章提出一种新的根据主题自动聚合短文本的主题模型，即伪文档主题模型（PTM）。和 SATM 一样，PTM 不依赖于额外信息就可以通过短文本自聚合增加词共现信息。和 SATM 不一样的地方是，PTM 的参数数量不随着数据增加，不易过拟合。此外，PTM 的时间复杂度远低于 SATM。PTM 的设计关键是引入伪文档，隐式的将短文本聚到伪文档中。通过这种方式，PTM 将在大量短文本上的主题学习转换到少量伪文档上，这增加了词共现信息，有利于提高主题模型学习的效果。本章还讨论了 PTM 的两个扩展模型，分别是有稀疏先验的 SPTM 和放宽了 PTM 模型假设的 EPTM。

通过在 4 个真实数据上的实验，我们验证了 PTM 可以在短文本上学习到高质量的主题。并且发现，当用 PTM 学习到的短文本主题表示训练分类器时，较少的训练数据

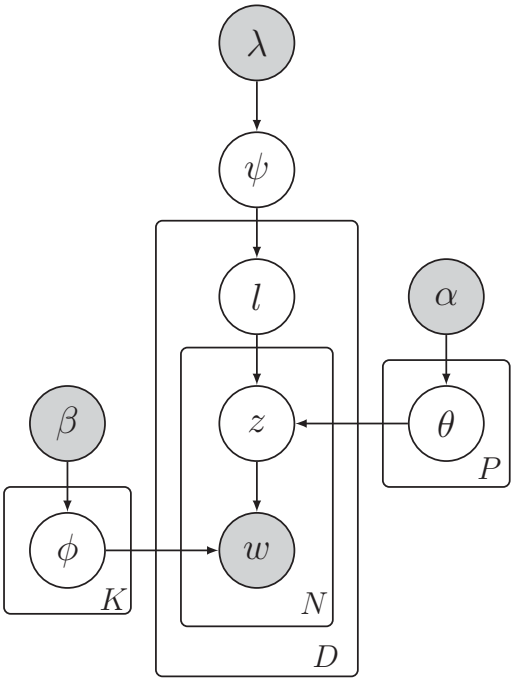


图 10　PTM 盘子表示法

就可以获得较好的分类精度。此外，还有两个有趣的发现。首先，当伪文档数量设置的较少时，SPTM 学习到的主题质量超过 PTM。其次，EPTM 表现不如 PTM 说明了 PTM 的一个短文本属于且只能属于一个伪文档的假设对于基于伪文档的短文本主题建模是很重要的，不应放宽该假设。

4.2　模型描述

本节，我们介绍可用于短文本的基于伪文档的主题模型（PTM）。PTM 假设大量短文本是从相对少量的潜在文档（latent document）中生成的。这里我们将潜在文档称为伪文档。学习伪文档对应的主题，而不是直接从短文本中学习主题，PTM 避免了短文本词稀疏等问题。合理的建模方式一方面使得 PTM 的参数数量不随着数据增加，降低了它过拟合数据的风险，另一方面保证了 PTM 学习算法的高效性。基于 PTM，我们还提出了带有稀疏先验的伪文档主题模型（SPTM）。SPTM 在伪文档-主题分布上采用了稀疏先验，这促使模型学习语义更明确的伪文档。

4.2.1　基础模型

现在我们给出 PTM 的形式化描述。假设有 K 个主题 $\{\phi_z\}_{z=1}^{K}$，每个主题是一个 V-维词典上的多项分布。一共有 D 篇短文本 $\{d_s\}_{s=1}^{D}$ 以及 P 篇伪文档 $\{d'_l\}_{l=1}^{P}$。短文本是已经观测到的文档，而伪文档是未观测到的潜在文档。我们引入一个多项分布 ψ 建模短文本

在伪文档上的分布。我们还假设一篇短文本属于且仅属于一篇伪文档。生成短文本中每个词的过程是从它所属伪文档对应的主题分布 θ 中采样一个主题 z，然后从中采样一个词 $w \sim \phi_z$。

图 10展示了 PTM 的盘子表示法（plate notation）。PTM 的完整生成过程如下：

1、 采样 $\psi \sim \mathrm{Dir}(\lambda)$

2、 为每一个主题 z:

（1） 采样 $\phi_z \sim \mathrm{Dir}(\beta)$

3、 为每一个伪文档 d'_l:

（1） 采样 $\theta_l \sim \mathrm{Dir}(\alpha)$

4、 为每一个短文本 d_s:

（1） 采样一个伪文档 $l \sim \mathrm{Multi}(\psi)$:

（2） 为每一个词 w_i in d_s:

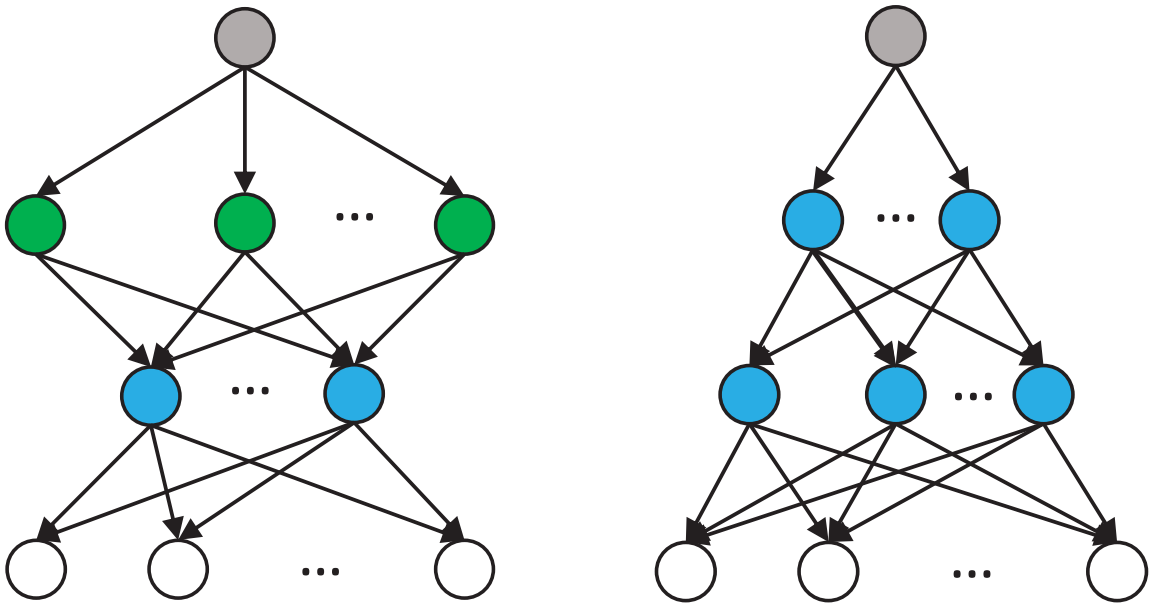
i. 采样一个主题 $z \sim \mathrm{Multi}(\theta_l)$:

ii. 采样第 i 个词 $w_i \sim \mathrm{Multi}(\phi_z)$

伪文档的引入是 PTM 能够克服短文本数据稀疏问题的关键。为了更好地理解这一点，下面我们做一些说明。假设一共有 D 篇短文本，每一篇平均有 N 个词。已经有理论证明，当 N 较小时，无论 D 多大，LDA [15] 都无法准确学习主题 [16]。这是由于短文本长度过短，词共现少导致的。然而，PTM 并不直接从 D 篇短文本中学习主题，而是通过 P 篇潜在的伪文档，通常 $P \ll D$。因此，我们可以大致的估算伪文档的平均词数 $N' = DN/P \gg N$。这意味着两点，首先伪文档是长文本，其次词共现增加了。简而言之，PTM 通过长的伪文档学习主题，避免了短文本数据稀疏和词共现缺少等问题。这确保了 PTM 可以在短文本上准确地学习主题。

4.2.2 模型比较

据我们所知，不依赖额外信息进行短文本自聚合的主题模型研究还很少见，目前只有 SATM [19]。尽管 PTM 和 SATM 解决短文本主题建模问题的思路一致，但是模型上有着根本的不同。二者的生成过程就完全不一样。SATM 假设短文本的生成分两阶段。第一阶段根据 LDA 的生成过程产生伪文档。第二阶段，按照一元混合模型 [22] 从伪文档中抽取短文本。第一阶段的设计导致 SATM 的统计推断算法采样每个词的隐变量的时间复杂度为 $O(PK)$，是 LDA 采样词主题时间复杂度的 P 倍。通常 P 取值成百上千，这意味着 SATM 的时间复杂度至少是 LDA 的几百倍。SATM 第二阶段的生成过程意味着不同短文本生成之间是相互独立的，导致 SATM 模型参数数量随着短文本数量的增加



(a) PTM　　(b) PAM

图 11　PTM 和 PAM 的模型对比示意图

而增加。这将导致严重的过拟合。与 SATM 不同，当给定短文本所属的伪文档后，PTM 生成短文本的过程和 LDA 一致，即采样词对应主题的时间复杂度均为 $O(K)$。而采样短文本所属伪文档的时间复杂度为 $O(P)$。因此，PTM 采样一个词的时间复杂度大致为 $O(P+K) \ll O(PK)$。此外，给定 P 和 K 之后，PTM 的参数数量为 $P+PK+KV$，并不随着短文本数量增加而增加。这避免了 PTM 过拟合。

和 PTM 相关的另一个模型是 PAM [93]。PAM 的提出是为了建模主题之间的关联关系。PAM 采用有向无环图（Directed Acyclic Graph 或 DAG）表示主题之间的关联关系，是比 LDA 表达能力更强的主题模型。特别的，Li 等人 [93] 在他们的工作中提出了一种特殊的 PAM 模型，采用了四层 DAG 的 PAM（如图 11(b)所示）。尽管这种 PAM 和 PTM 在结构上类似（如图 11所示），但是他们本质上是不同。PAM 的第二层结点代表超级主题（super=topics），而第三层结点代表子主题（sub-topics）。两类主题均表示为蓝色结点。超级主题建模的是子主题之间的语义关系，而子主题建模的是词（图 11中的白色结点）之间的语义关系。因此，通常超级主题的数量远小于子主题的数量。然而，PTM 的第二层结点（图 11(a)中的绿色结点）代表伪文档，其数量远大于第三层的主题结点。因为伪文档的作用是将主题分布相似的短文本合在一起。此外，PTM 和 PAM 的一个关键区别在于 PTM 约束每篇短文本只属于一篇伪文档，而 PAM 并没有这个假设。后面我们对 PTM 做了相应的扩展，即 EPTM。届时会说明取消了每篇短文本只属于一篇伪文档的假设可能导致模型性能下降。

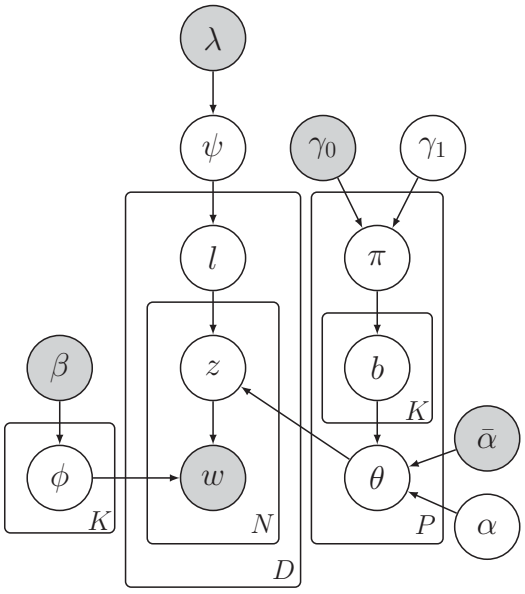


图 12　SPTM 盘子表示法

4.2.3　稀疏模型

如上文所述，伪文档的作用是将主题分布相似的短文本合在一起。因此，一个自然的猜测是当伪文档数目设置比较少的时候，它的主题可能比较混杂。为了应对这种情况，我们为 PTM 的伪文档主题分布引入一种 “Spike-and-slab” 稀疏先验，提出了 SPTM 模型。

“Spike-and-slab” [94] 是一种数学上较为完善的贝叶斯参数选择方法。它可以有效地将分布的平滑性和稀疏性解耦。具体来说，稀疏先验中引入的 Bernoulli 分布可以表示特定变量 “开” 和 “关” 的状态。因此，模型有了决定一个变量是不是起作用的能力。在 SPTM 中，一个主题是不是出现在某篇伪文档中由稀疏先验决定。而不再是 PTM 的所有主题都可以在所有伪文档中出现。

值得注意的是，“Spike-and-slab” 稀疏先验可能导致病态分布。例如，当所有随机变量都处于 “关” 的状态时，联合分布是无意义的。Wang 和 Blei [95] 为分布引入一个虚拟变量，通过约定该变量一定是 “开” 的状态解决了上述问题。然而该方法会导致模型的统计推断变得困难。为了避免模型求解变得复杂，我们采用 Lin 等人 [24] 提出的方法。具体而言，当一个变量被关闭时，我们给它一个非常弱的先验。相反的，变量开启时，给它正常的先验。通过不完全关闭一个变量，避免了所有变量均被关闭导致的病态分布问题。为了更好的理解 SPTM，我们首先给出主题选取器（topic selectors）以及平滑先验（smoothing prior）、弱平滑先验（weak smoothing prior）的定义。

定义 4.1： 对于伪文档 d'_l 而言，一个主题选取器 $b_{l,k}$ 是一个二进制变量，它指示了主题

k 是不是和 d'_l 有关。$b_{l,k}$ 是采样自 d'_l 对应的参数为 π_l 的 *Bernoulli* 分布。

定义 4.2： 平滑先验是指用来平滑被主题选取器选中的主题对应的 *Dirichlet* 分布超参数 α。弱平滑先验是指用来平滑没有被主题选取器选中的主题对应的 *Dirichlet* 分布超参数 $\bar{\alpha}$。因为 $\bar{\alpha} \ll \alpha$，所以超参数 $\bar{\alpha}$ 被称为弱平滑先验。

主题选取器又被称为 “Spikes”，而平滑先验或者弱平滑先验又叫作 “Slabs”。通过引入主题选取器和平滑（弱平滑）先验，我们成功解耦了伪文档主题分布的稀疏性和平滑性。给定主题选取器 $\vec{b_l} = \{b_{l,k}\}_{k=0}^{K}$，伪文档 d'_l 的主题分布 θ 从 $\mathrm{Dir}(\alpha\vec{b_l} + \bar{\alpha}\vec{1})$ 分布采样获得。弱平滑先验 $\bar{\alpha}$ 的引入解决了 “Spike-and-slab” 先验可能导致的病态分布问题，同时保证了分布的稀疏性。

图 12给出了 SPTM 的盘子表示法。SPTM 对应的伪文档生成过程如下：

1、 为每一个伪文档 d'_l:

（1） 采样 $\pi_l \sim \mathrm{Beta}(\gamma_0, \gamma_1)$

（2） 为每一个主题 z:

i. 采样主题选取器 $b_{l,k} \sim \mathrm{Bernoulli}(\pi_l)$,
$\vec{b_l} = \{b_{l,k}\}_{k=0}^{K}$.

（3） 采样 $\theta_l \sim \mathrm{Dir}(\alpha\vec{b_l} + \bar{\alpha}\vec{1})$

4.3 统计推断

伪文档主题模型准确的后验概率是无法计算的，我们只能通过 collapsed Gibbs 采样近似估计。Collapsed Gibbs 采样有容易推导采样公式，编写代码简单，运行相对高效和能够估计全局最优解等优势。下面我们首先给出 SPTM 的采样公式，然后在本节的最后介绍如果根据 PTM 的采样。

积分积掉 θ、ϕ、ψ 和 π，需要采样的隐变量为短文本对应的伪文档 l，每个词对应的主题 z 以及伪文档对应主题分布先验中的主题选取器 b。Dirichlet 超参数 α，Beta 分布超参数 γ_1 也是通过采样得到，而 $\bar{\alpha}$ 设置为 10^{-7}，γ_0 设置为 1.

采样伪文档 l。给定其他隐变量，采样 l 和 Dirichlet Multinomial mixtures 模型 [96] 的过程相似：

$$\begin{aligned} p(l_{d_s} = l|rest) &\propto \frac{M_{l,\neg d_s}+\lambda}{D-1+P\lambda} \frac{\frac{\prod_{z\in d_s}\Gamma\left(N_l^z+b_{l,z}\alpha+\bar{\alpha}\right)}{\prod_{z\in d_s}\Gamma\left(N_{l,\neg d_s}^z+b_{l,z}\alpha+\bar{\alpha}\right)}}{\prod_{i=1}^{N_{d_s}}(N_{l,\neg d_s}+|A_l|\alpha+K\bar{\alpha}+i-1)} \\ &= \frac{M_{l,\neg d_s}+\lambda}{D-1+P\lambda} \frac{\prod_{z\in d_s}\prod_{j=1}^{N_{d_s}^z}\left(N_{l,\neg d_s}^z+b_{l,z}\alpha+\bar{\alpha}+j-1\right)}{\prod_{i=1}^{N_{d_s}}(N_{l,\neg d_s}+|A_l|\alpha+K\bar{\alpha}+i-1)}, \end{aligned} \tag{4.1}$$

其中 M_l 归到第 l 个伪文档 d'_l 下的短文本数量。N_{d_s} 是第 s 个短文本 d_s 的长度，而 $N^z_{d_s}$ 是 d_s 中被分配到主题 z 下的词数。N^z_l 是伪文档 d'_l 中被分配到主题 z 下的词数，而 N_l 是 d'_l 一共有多少词。所有带有 $\neg d_s$ 下标的计数都意味没有计入 d_s 相应计数的结果。$b_{l,z}$ 是伪文档 d'_l 上主题 z 对应的选取器。$A_l = \{z : b_{l,z} = 1, z \in \{1, \cdots, K\}\}$ 是 $\vec{b_l}$ 等于 1 的下标的集合，而 $|A_l|$ 是 A_l 的基数。

采样词主题 z。词主题 z 的采样和 LDA Gibbs 采样类似 [42]。区别在于 θ 不再是短文本对应的主题分布而是伪文档的，并且 θ 是从 "Spike-and-slab" 稀疏先验采样得到。具体采样公式如下：

$$p(z_{d_s,i} = z|rest) \propto (N^z_{l_{d_s}} + b_{l_{d_s},z}\alpha + \bar{\alpha}) \frac{N^{w_{d_s,i}}_z + \beta}{N_z + V\beta}, \tag{4.2}$$

其中 N^w_z 是词 w 被分配到主题 z 下的词数，而 $N_z = \sum_{w=0}^{V} N^w_z$。

采样主题选取器 b。为了采样 $\vec{b_l}$，我们用 π_l 作为辅助变量。定义 $B_l \triangleq \{z : N^z_l > 0, z \in \{1, \cdots, K\}\}$ 为伪文档 d'_l 中被分配的主题集合。π_l 和 $\vec{b_l}$ 的联合条件分布为：

$$p(\pi_l, \vec{b_l}|rest) \propto \prod_z p(b_{l,z}|\pi_l) p(\pi_l|\gamma_0, \gamma_1) \frac{I[B_l \in A_l]\Gamma(|A_l|\alpha + K\bar{\alpha})}{\Gamma(N_l + |A_l|\alpha + K\bar{\alpha})}, \tag{4.3}$$

其中 $I[\cdot]$ 为指示函数。有了上面的联合条件分布，我们可以交替地根据 π_l 采样 $\vec{b_l}$ 以及根据 $\vec{b_l}$ 采样 π_l，最终得到 $\vec{b_l}$ 的结果。值得注意的是，Wang 等人 [95] 为了采样更快收敛，他们积分掉 b 后采样 π。不过那是因为词典大小 V 很大，导致最优解的搜索空间很大。在 SPTM 中，由于 K 比 V 小很多，因此我们积分掉 π 后采样 b。

对于超参数 α, 我们用以对称高斯分布（symmetric Gaussian）为提议分布（proposal distribution）的 Metropolis-Hastings 方法采样。对于集中参数（concentration parameter）γ_1，我们采用 Teh 等人 [62] 研发的为 Gamma 先验采样超参数的方法。

目前为止，我们已经介绍了 SPTM 的 collapsed Gibbs 采样算法。现在我们简单描述一下 PTM 对应的算法。积分掉 θ、ϕ 和 ψ，PTM 中需要采样的隐变量为短文本对应的伪文档 l 以及词主题 z。通过将公式 5.1中的 $b_{l,z}\alpha + \bar{\alpha}$ 和 $|A_l|\alpha + K\bar{\alpha}$ 分别替换为 α 和 $K\alpha$，我们就得到了 PTM 采样 l 的公式。类似的，通过将公式 4.2中的 $b_{l,z}\alpha + \bar{\alpha}$ 替换为 α，我们就得到了 PTM 采样 z 的公式。

PTM 中采样每篇短文本对应的伪文档编号 l 的时间复杂度为 $O(P)$，采样每个词对应主题 z 的时间复杂度为 $O(K)$。因此，PTM 时间复杂度平均到每个词的采样过程上大致为 $O(P + K)$。由于 SPTM 需要采样主题选取器 b，因此它的时间复杂度略大于 PTM。

PTM 和 SPTM 学习的是伪文档 d'_l 的主题分布 θ_l。为了获得短文本 d_s 的主题分布 θ_s,

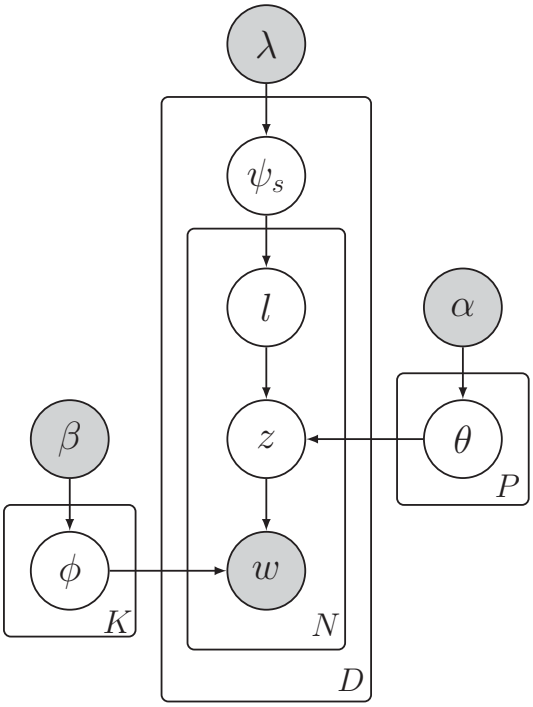


图 13　SPTM 盘子表示法

我们采用经验估计。有了 d_s 中每个词对应的主题，即 $\vec{z}$，我们可以得到 $\theta_{s,z} = \frac{N^z_{d_s}+\alpha}{N_{d_s}+K\alpha}$。相应的，对于 SPTM 我们可以得到 $\theta_{s,z} = \frac{N^z_{d_s}+b_{d_s,z}\alpha+\bar{\alpha}}{N_{d_s}+|A_{d_s}|\alpha+K\bar{\alpha}}$。

4.4　模型扩展与讨论

正如前文所述，PTM 限制每篇短文本属于且只属于一篇伪文档。本节，我们将修改 PTM 模型以取消该限制，并讨论这么做带来的影响。

PTM 通过为每篇短文本采样伪文档编号 $l \sim \psi$ 将其分配到对应伪文档中。$\psi \sim Dir(\lambda)$ 是短文本在伪文档上的多项分布。通过假设每篇短文本各自有一个在伪文档上的多项分布 ψ_s，我们便可以取消 PTM 的限制，进而得到增强的 PTM（EPTM）。EPTM 的盘子表示法展示在图 13中。

之所以称之为 EPTM，是由于它比 PTM 更加灵活。因为短文本中不同的词可以分配到不同的伪文档中。尽管 EPTM 比 PTM 更灵活和普适，但是它建模短文本的实际效果可能不如 PTM 或者 SPTM。如果我们假设图 13中的 $\vec{z}$ 是观测到的变量，那么 l 的生成过程就和 LDA 中生成 z 的过程一致了。那么 Tang 等人 [16] 讨论的 N 很小时 LDA 无法学习到准确的 ϕ 的情况就可以应用到这里。即，N 很小时，EPTM 无法准确的学习图 13中的 θ。这必然会限制 EPTM 在短文本上的性能。

除了理论上 EPTM 并不适用于建模短文本，它的训练过程也变的非常耗时。因为为

表 5　数据集的统计指标

Data set	# Documents	Vocabulary size	Avg. document length
News	29,200	11,007	12.4
DBLP	55,290	7,525	6.4
Questions	142,690	26,470	4.6
Tweets	182,671	21,480	8.5

每个词采样 (l, z) 的时间复杂度为 $O(PK)$。实践中为了减少采样的时间复杂度，我们采用 alias sampling 技术 [49]，但是采样算法需要消耗大量内存来保存加速采样过程的 alias table。

4.5　实验结果与分析

4.5.1　实验设置

数据集

为了测试伪文档主题的性能，我们找了四个真实的短文本数据集。表 5给出了它们的一些统计结果。下面分别对它们作简单介绍。

News： 该数据集[1]包含 29200 条英文新闻。每条新闻属于一个类别。所有类别包括 sport, business, U.S., health, sci&tech, world 和 entertainment。我们仅使用新闻的摘要，因为它是典型的短文本。

DBLP： 我们收集了计算机学科 6 个研究领域会议论文的标题。具体的领域包括：数据挖掘、机器视觉、数据库、信息检索、自然语言处理和机器学习。最终获得了 55290 个短文本，每个都标注了 6 个研究领域中的一个。

Questions： 这个数据集是 Yan 等人 [23] 从百度知道[2]抓取的 142690 条问题。每个问题属于 35 个类别之一。

Tweets： Zubiaga 等人 [97] 抓取并标注了大量的 tweets。他们抓取了包含 URL 的 tweets，并且用 URL 指向的网页的类别作为对应 tweet 的类别。网页的类别是根据开放目录计划（Open Directory Project 或 ODP）定义的。该数据集包含 10 个不同类别，一共 360k 左右的 tweets。我们从中选择了 9 个话题相关的类别，并从中抽样了 182671 条 tweets。

[1] http://acube.di.unipi.it/tmn-dataset/

[2] http://zhidao.baidu.com

基准方法

Latent Dirichlet Allocation (LDA)。作为最经典的主题模型，LDA [15] 可以通过将 Dirichlet 参数设置的趋近于零来保证模型的稀疏性。这里 LDA 的实现使用的是开源的 jGibbLDA[3]。

Mixture of Unigrams (MU)。MU [22] 最大的特点是假设每个文档所有的词都只属于一个主题。MU 通过该假设强制每篇文档的主题表示达到最大的稀疏度。对于长文本，该假设也许不太合理，但是在短文本上也许是可行的。

Dual Sparse Topic Model (DSTM)。DSTM [24] 是近期提出的带有稀疏先验的主题模型。它将 LDA 的文档-主题分布以及主题-词分布的 Dirichlet 先验均修改为 “Spike-and-Slab” 稀疏先验。这使得 DSTM 可以适用于短文本一类的稀疏数据。

Self-aggregate Topic Model (SATM)。和 PTM 一样，SATM [19] 也能够在短文本自聚合构成的伪文档上学习主题。但是，因为它的参数数量随着数据增加而增加，所以容易过拟合。通过调整训练数据的规模，对比 PTM 和 SATM 的结果可以说明该问题。

评价指标

主题一致性（Topic Coherence）。主题模型的评价方式仍然是一个开放问题。通常使用的混淆度（perplexity）指标被证明和主题的可读性并不是很相关。这意味着模型混淆度低并不意味着学到了更容易理解的主题。而且，很多短文本主题模型并不直接从短文本学习主题，例如 SATM 和 PTM，导致混淆度不再是一个具有普适意义的主题模型评价指标。因此很多方法转而使用一致性指标评价主题。主题一致性已经被证明和主题的可读性是相关的。此外，最近有研究 [19] 指出 UMass 一致性指标 [58] 并不适用于短文本，并推荐 UCI 主题一致性 [57] 评价短文本主题模型。UCI 主题一致性的技术需要利用足够大且足够通用的外部语料。Wikipedia 是常用的外部语料。然而它可能适用于评价经过良好编辑的诸如新闻或论文等文本，却不一定适用于用户产生的诸如 tweets 等文本。

基于上述考虑，我们使用 UCI 一致性指标评价 News 和 DBLP 的结果。UCI 一致性指标使用点互信息（PMI）计算主题的一致性。给定一个主题 z，我们选择它的 top-N 主题词 $w_1, w_2, \cdots, w_N$，然后计算这些词构成的每个词对的 PMI 的均值：

$$\mathrm{PMI}(z) = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{4.4}$$

[3] http://jgibblda.sourceforge.net

表 6　五折交叉验证的短文本分类结果

	News			DBLP			Questions			Tweets		
	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
PTM	*0.755*	*0.757*	*0.754*	**0.667**	**0.672**	**0.668**	**0.532**	**0.554**	**0.536**	0.561	*0.568*	*0.559*
SPTM	**0.760**	**0.761**	**0.759**	*0.661*	*0.667*	*0.663*	*0.530*	*0.552*	*0.532*	0.551	0.558	0.550
SATM	0.697	0.702	0.686	0.657	0.662	0.654	0.312	0.353	0.297	*0.599*	**0.605**	**0.594**
LDA	0.727	0.732	0.728	0.613	0.624	0.614	0.502	0.529	0.506	0.553	0.560	0.546
DSTM	0.720	0.724	0.720	0.619	0.628	0.620	0.489	0.515	0.491	0.539	0.547	0.535
MU	0.697	0.617	0.626	0.640	0.643	0.638	0.511	0.526	0.509	**0.634**	0.546	0.546

其中 $p(w_i, w_j)$ 是词对 (w_i, w_j) 出现在同一个滑动窗口的联合概率，$p(w_i)$ 是词 w_i 出现在一个滑动窗口的边缘概率。这些概率值是在 Wikipedia 的文档上统计得到的。我们使用所有主题的一致性指标的均值作为一个主题模型的主题一致性结果。实验中 N 的值统一设置为 10。

分类指标。主题模型一个常用的外部评价方式是文档分类。因此，我们使用主题模型学到的短文本语义表示去训练分类器，然后使用宏平均的精确度、召回率和 F-值评价分类效果。

参数设置

实验中的所有模型的主题数量均设置为 100。除非有特殊说明，SATM 和我们方法的伪文档数量统一设置为 1000。除了 SPTM 和 DSTM，所有方法的 Gibbs 采样均执行 2000 轮迭代。由于 SPTM 和 DSTM 需要采样稀疏先验中的 2 值随机变量，它们 Gibbs 采样均执行 3000 轮迭代，以确保模型的收敛。

对于 LDA，我们设置 $\alpha = 0.1$，$\beta = 0.01$。因为弱先验的 LDA 在短文本上表现更好。类似的，我们设置 MU 的 $\alpha = 0.1$, $\beta = 0.01$。我们设置 DSTM 的 $\pi = 0.1$, $\gamma = 0.01$。该设置和 DSTM 原论文中的不一致，但是该设置下的 DSTM 表现更好。DSTM 中 $\bar{\pi}$ 和 $\bar{\gamma}$ 的设置和原论文一致，均取值 10^{-12}。SATM 的参数设置和原论文一致。对于 PTM 和 EPTM，我们设置 $\alpha = 0.1, \lambda = 0.1$ 以及 $\beta = 0.01$。对于 SPTM，我们设置 $\gamma_0 = 0.1$，$\bar{\alpha} = 10^{-12}$。所有报告的实验结果均是 5 次训练结果的均值。

4.5.2　短文本分类

首先我们对比所有主题模型在文档分类任务上的表现。这里，主题模型被用作降维工具将短文本表示为主题分布。

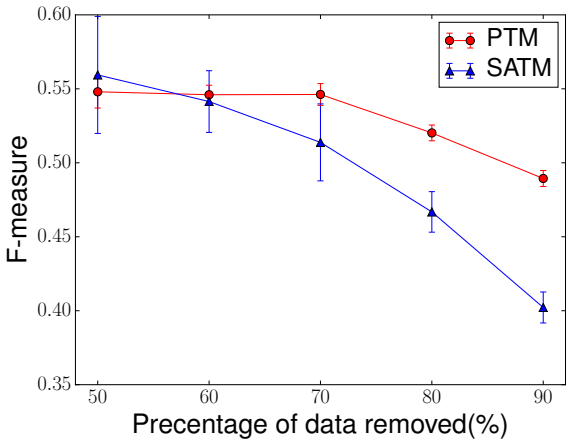


图 14　SATM 过拟合的图示

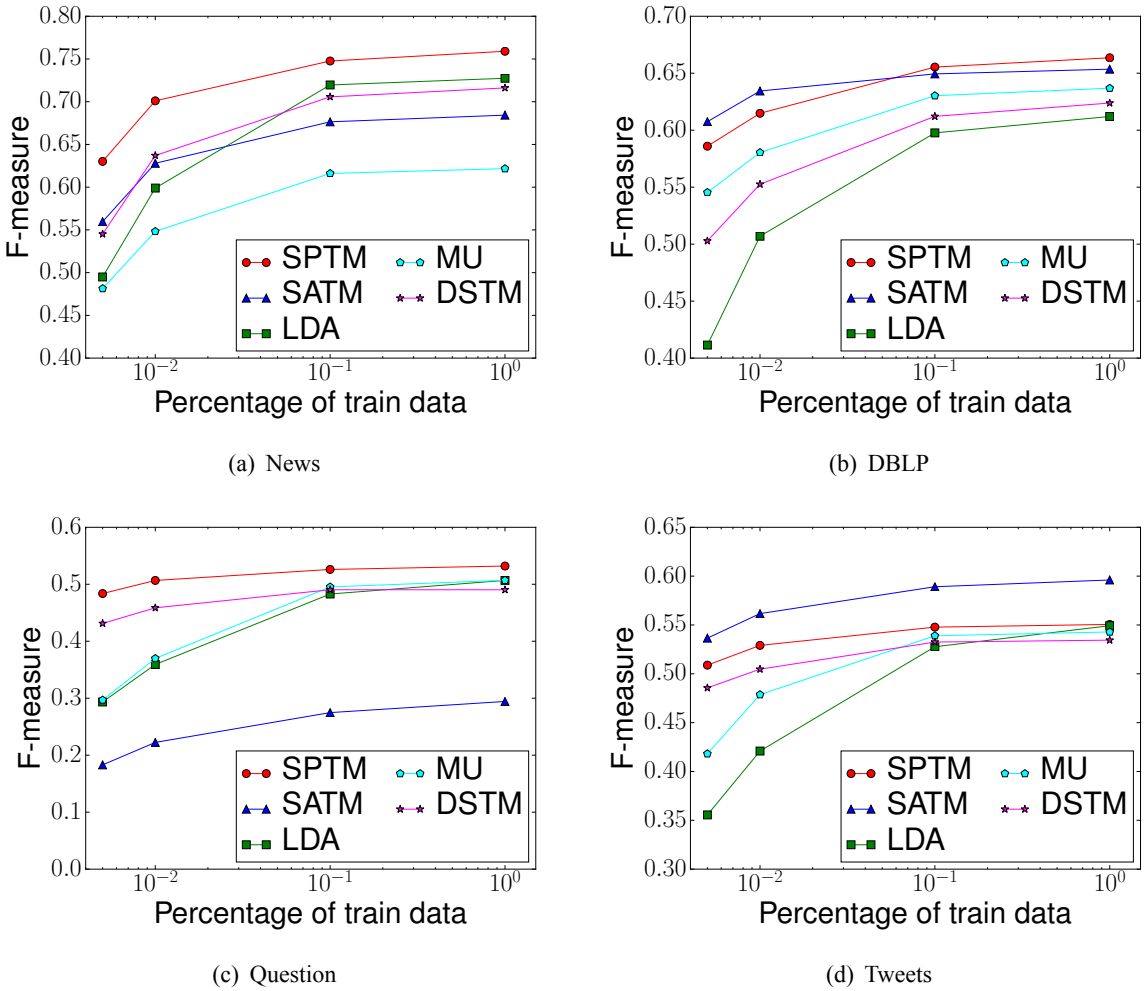
五折交叉验证的分类结果。每个主题模型训练得到所有短文本的主题表示后进行五折交叉验证。分类器采用 LIBLINEAR[4]。表 6给出了宏平均的准确度、召回率以及 F-值。最好的结果用黑体表示，次好的结果用斜体表示。

从结果中可以发现，在 News、DBLP 以及 Question 这三个数据集上，最好的以及次好的结果均由 PTM 或 SPTM 取得。这表明了伪文档主题模型在学习短文本主题表示比基准方法拥有更好的表现。特别的，我们的方法比 LDA 有明显优势。这说明了短文本聚合为长的伪文档有利于主题模型更准确地学习主题，进而得到更好的短文本主题表示。我们的方法在所有数据集上均超过了 DSTM 和 MU，这说明了短文本聚合比为模型增加稀疏性更可靠。

一个有趣的现象是 SATM 在 Tweets 上取得了最好的效果，但是在其他三个数据集上却效果平平。为了分析 SATM 表现不稳定的原因，我们统计了四个数据集每个类别的平均文档数目。News、DBLP、Questions 和 Tweets 的统计结果分别为 4171、9215、4077 和 20297。不难发现，SATM 的性能似乎依赖于训练数据量。这和前文关于 SATM 容易过拟合的分析结论一致。为了保证 SATM 有好的表现，需要给它足够的训练数据。即便训练数据容易获得，但是 SATM 的训练非常耗时（时间复杂度为 $O(PK)$）。这限制了 SATM 在大规模数据上的应用。例如，SATM 在 Tweets 上的训练用了 7 天的时间，然而 PTM 和 SPTM 均在一天之内完成了训练。值得注意的是，SATM 的实验已经采用了 sparse Gibbs 采样技术 [48] 做了加速。如果采用普通的 Gibbs 采样，SATM 在 Tweets 上的训练大致需要花 1 个月的时间！

为了进一步验证 SATM 确实容易过拟合，我们在 Tweets 设计了一个专门的实验。从原始的 Tweets 数据集从随机抽取掉 50% 至 90%，然后在修改后的数据集上分别训练 PTM 和 SATM。图 14给出了五折交叉验证的宏平均 F-值。从结果中可以发现，随着

[4]http://www.csie.ntu.edu.tw/ cjlin/liblinear/



(a) News (b) DBLP

(c) Question (d) Tweets

图 15　调整训练集大小的分类结果

tweets 数量的减少，SATM 的性能下降很快，而 PTM 则很稳定。这个结果直接验证了上面的讨论，SATM 很容易过拟合。当训练数据有限时，它的性能很容易变差。

调整分类器训练集大小的结果。 当训练样本很少时，用主题表示的训练文档可以给分类器更好的泛化能力 [59]。为了测试哪个主题模型学习到的表示可以为分类器提供更好的泛化能力，我们设计了该实验。在所有数据集上，我们用 80% 的短文本做训练，剩下的 20% 做测试。分类器依旧使用 LIBLINEAR。从训练集中采样 0.5% 至 100% 的短文本分别训练分类器并在所有测试短文本上计算 F-值。

图 15给出了不同采样比例下的分类结果。为了图示更加清晰，我们省去了 PTM 的结果，因为它的表现和图中的 SPTM 非常接近。从结果中不难发现，SPTM 在所有数据集上均能够超过 LDA、MU 和 DSTM。这很好地说明了伪文档主题模型的鲁棒性。SATM 在 Tweets 上超过了 SPTM，然而，这个结果是由于它过拟合导致的。

表 7　New 和 DBLP 上的 UCI 主题一致性结果

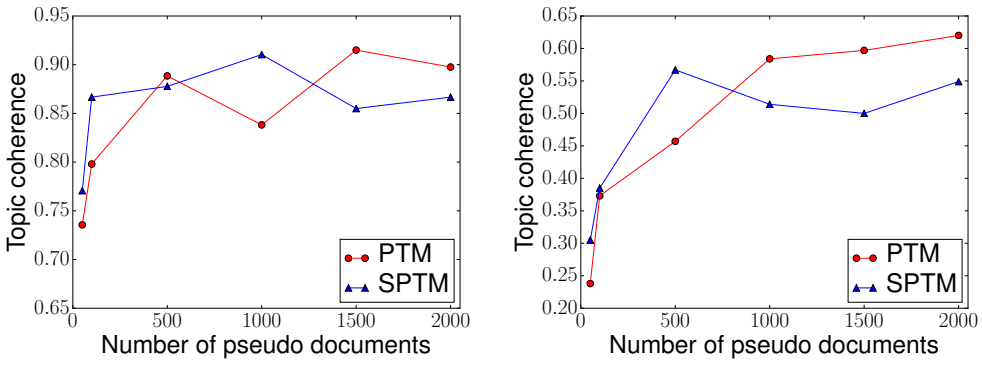
	PTM	SPTM	SATM	LDA	MU	DSTM
News	0.838	**0.910**	0.187	0.795	0.391	0.693
DBLP	**0.584**	0.514	-1.933	0.548	0.525	0.389

4.5.3　主题一致性

表 7给出了 PTM、SPTM 以及所有基准方法在 News 和 DBLP 上的 UCI 主题一致性结果。如前文所述，UMass 主题一致性并不适于短文本 [19]。UCI 主题一致性更适合，但是它需要外部语料计算词对之间共现的联合概率。对于 News 和 DBLP，合适的外部语料比较容易获得。这里我们使用 Wikipedia。然而对于 Tweets 和 Questions 并没有合适的外部语料，因此我们不在这两个数据集上对比 UCI 主题一致性。

从表中的结果可以发现 SPTM 在 News 上取得了最好的效果，而 PTM 在 DBLP 上取得了最好的效果。伪文档主题模型的主题一致性超过 LDA 再次印证了从短文本聚合成的伪文档上学习主题有利于学习更好的主题。在 News 和 DBLP 上，LDA 的表现是基准方法中最好的，而 SATM 的表现最差。比较有意思的一个结果是带有弱先验参数的 LDA 表现出色，甚至超过了 DSTM。尽管 DSTM 的稀疏先验在理论上更完备，然而实际表现却不如带有弱先验参数的 LDA。这可能是 DSTM 的模型复杂度（参数数量）高于 LDA，导致它在稀疏数据上的训练面临更大的挑战。MU 在 News 上表现不佳，但是在 DBLP 取得了不错的效果。这可能是由于 News 的标题倾向于包含不止一个主题，导致 MU 的假设在 News 引入了不少的错误。由于过拟合的缘故，SATM 需要一个更大的数据集来训练数据。这导致它在 News 和 DBLP 很难学到一致性好的主题。

4.5.4　参数敏感性



(a) News　　(b) DBLP

图 16　主题一致性随伪文档数量的变化

表 8　PTM、SPTM 和 EPTM 的主题一致性结果

	PTM	SPTM	EPTM
News	0.838	0.910	0.780
DBLP	0.584	0.514	0.489

根据 Tang 等人的工作 [16] 可知，文档平均长度 N 以及文档数量 D 是影响 LDA 学习准确性的关键。当 D 或者 N 很小时，主题模型无法从文档集中准确学习主题。伪文档主题模型从 P 个伪文档中学习主题，而 P 的设置直接关系到伪文档的数量，间接决定了伪文档的平均长度。因此 P 的设置是伪文档主题模型准确建模短文本的关键。直觉上理解，P 设置的过小模型将难以学到一致性好的主题。

类似于 Tang 等人的做法，我们也通过调整 P 来观察主题一致性的变化，据此讨论 P 伪文档主题模型学习准确性的影响。具体的，我们将 P 的值从 50 调整到 2000，观察 PTM、SPTM 在 News 和 DBLP 上 UCI 主题一致性的变化。图 16展示了该实验的结果。

从结果中，我们观察到了两点。第一点，P 设置过小时，伪文档主题模型学习到的主题一致性较差。如图 16(a)所示，在 News 上，PTM 在 P 等于 50 和 100 时主题一致性较差。而当 $P \geq 500$ 时，主题一致性较好。该结果和理论分析是一致的。因为主题模型需要足够多的文档来确保训练过程的准确性。第二点，带有稀疏先验的伪文档主题模型在 P 设置较小时表现的更好。如图 16(a)所示，当 $P = 100$ 时，SPTM 的主题一致性好过 PTM。类似现象在图 16(b)中也有，当 $P = 500$ 时，SPTM 的主题一致性也明显好过 PTM。当 P 设置较小时，伪文档的主题可能比较混杂，而稀疏先验可以帮助它学习更相关的主题，因为稀疏先验可以消除伪文档和主题之间一些不必要的关联关系。因此 SPTM 在 P 设置较小时比 PTM 更鲁棒。

另一个有趣的现象是当 P 设置的足够大时，PTM 总是比 SPTM 的效果好。当 P 足够大后，稀疏先验引入的额外随机变量给模型的训练带来了困难。这也许是当 P 足够大时 SPTM 表现不如 PTM 的原因。这个结果也间接说明了 DSTM 在实践中效果不如 LDA 的原因。

表 9　PTM、SPTM 和 EPTM 的分类结果

	News			DBLP		
	precision	recall	f-measure	precision	recall	f-measure
PTM	0.755	0.757	0.754	0.667	0.672	0.668
SPTM	0.760	0.761	0.759	0.661	0.667	0.663
EPTM	0.749	0.751	0.749	0.645	0.654	0.647

4.5.5　扩展模型的验证

PTM 和 SPTM 均假设每篇短文本属于且仅属于一篇伪文档，而 EPTM 则取消了这个假设。虽然 EPTM 模型上更为灵活，但是根据前文的讨论，它的性能应该不如 PTM 和 SPTM。为了验证这一点，我们分别在 UCI 主题一致性和五折交叉分类上对比了 EPTM 和 PTM 以及 SPTM。表 8给出了三个模型主题一致性的结果。可以发现 EPTM 学习到的主题一致性确实差于另外两个方法。表 9给出了三个模型分类的结果。同样的，EPTM 的效果又是最差的。两组结果均验证了我们的讨论，也说明每篇短文本属于且仅属于一篇伪文案的假设是必要的。

4.6　小结

本章中，我们提出了伪文档主题模型（PTM）。通过将大量短文本聚合为少量的伪文档，PTM 增加了词共现信息，使其可以更准确地从短文本中学习主题。我们还提出带有稀疏先验的伪文档主题模型（SPTM）。当伪文档数目设置较少时，SPTM 可以提升 PTM 的效果。此外，我们还验证了一个短文本属于且仅属于一个伪文档的假设是必要的。在 4 个真实数据集上的大量实验结果说明了 PTM 和基准方法相比是有优势的。

本章的研究成果已于 2016 年被数据挖掘顶级会议 SIGKDD 录用，论文题目为“Topic Modeling of Short Texts: A Pseudo-Document View”。

第五章　伪文档 N-gram 主题模型

5.1　引言

由于缺少文档级词共现信息，经典主题模型 Latent Dirichlet Allocation（LDA）[15] 在短文本上表现不好已经变的众所周知。近年来，针对短文本的主题模型的研究收到越来越多的关注 [19, 21, 23, 24, 56]。这些工作的一个潜在缺陷是它们均沿用了经典主题模型采用的词袋模型的假设。该假设可以提高模型训练的效率，但是忽视词序信息可能严重影响短文本的主题建模。下面我们给出两个详细的理由：

- 拥有相同词袋表示的两个句子可能表达了完全不同的意思。例如，“the department chair couches offer” 和 “the chair department offers couche” 有着相同的一元统计信息，但是表达的主题却完全不同 [25]。和长文本不同，许多短文本只包含单个句子。这导致词序对于短文本主题建模而言是一个比较重要的难以忽略的信息。
- 一个词语搭配（collocation）或者短语（phrase）表示的意思有时并不是它包含的词的含义的简单组合。例如，*power supply* 指的是电源设备，而 *power* 或 *supply* 单独来看和电源关联不大。因为 LDA 容易将 collocation 中的词分配到不同主题下 [98]，所以 *power* 和 *supply* 可能被不同主题生成。例如，*power* 有电力相关主题生成，而 *supply* 由经济类主题生成。

上面的情况说明了词序对短文本主题建模的重要性。现有的工作通常采用 collocation（phrase 或者较短的相邻词序列）将词序信息引入主题模型 [25, 98–100]。将相邻的词拼接为 collocation 的方法减少了词共现信息。这对长文本的主题建模影响不大，但是会导致短文本的内容稀疏问题更加严重。此时，不难发现一个矛盾。一方面，词序信息对于短文本主题建模很重要。另一方面，现有的手段会导致短文本内容稀疏问题更加严重。这个矛盾促使我们设计一个新的模型，它既能够引入短文本上的词序信息，又能兼顾内容稀疏的问题。

本文中，我们提出伪文档 N-gram 主题模型（PTNG）。PTNG 通过将短文本聚合为长文本解决内容稀疏问题。此外，PTNG 可以根据上下文自动学习单词或者 collocations，并为它们分配主题。因此，PTNG 可以更准确地从短文本中学习主题。据我们所知，本文是第一个在短文本主题建模过程中考虑词序信息的工作。在三个实际数据集上的实验表明，PTNG 可以学习高质量的主题，而且学到的短文本主题表示在分类任务中也有很好的表现。

本章剩余部分组织如下：5.2节介绍了 N-gram 主题模型的相关工作，为了避免重复省去了短文本主题模型相关工作的介绍；5.3节介绍了 PTNG 模型的生成过程和统计推断方法；5.4节给出了实验结果和分析；5.5节给出了本章小结。

5.2　N-gram 主题模型相关工作

大部分主题模型采用词袋模型的假设，即词和词之间的生成过程相互独立。虽然词袋模型可以使模型训练过程更高效，但是忽视了词序相关的有用信息。为了将词序信息引入主题模型，将相邻词拼接为 collocations 并看作一个整体是最常用的手段。之前基于 collocation 的工作可以大致分为两类。一类采用流水线的方式，首先挖掘短语，然后训练主题模型 [101–103]。另一类方法是基于 LDA 做模型扩展 [25, 98–100]。流水线的方法需要挖掘短语的预处理过程，这和语言以及领域有关。因此，本文关注的是基于 LDA 的扩展方法。

Wallach [25] 提出了 bigram 主题模型。该模型假设主题是 biram 的多项分布，而不是 unigram 的分布。一个词的生成由前一个词和自己的主题共同决定。这么做的缺点在于每一对连续的词均被看作 bigram，然而实际情况是大部分词是 unigram。Griffiths 等人 [100] 提出的 LDACOL 可以学习两个相邻词是否构成 bigram。通过为每个词的生成过程引入了一个名为 bigram 状态的二值随机变量，LDACOL 可以学习两个相邻词是否构成 bigram。Wang 等人 [99] 对 LDACOL 做了改进提出了 TNG。在 LDACOL 的生成过程中，bigram 第二个词的生成仅和前一个词相关。而在 TNG 中，bigram 第二个词的生成还和自己的主题有关。TNG 的一个潜在缺陷是它并不约束 n-gram 中的词来自同一个主题。该问题在 Lindsey 等人 [98] 提出的 PDLDA 中得到解决。

5.3　模型描述与统计推断

本节我们首先具体描述 PTNG 模型，然后给出关于 PTNG 的一些讨论，最后给出模型的统计推断方法。PTNG 假设大量短文本是通过少量伪文档生成的，此外 PTNG 能够根据上下文和主题自动探测 collocations。

5.3.1　模型描述

现在我们给出 PTNG 的正式描述。假设有 D 个观测到的短文本 $\{d_s\}_{s=1}^{D}$ 和 P 个潜在的伪文档 $\{d'_l\}_{l=1}^{P}$。我们引入 ψ 表示短文本在伪文档上的多项分布，还约束每个短文本属于且仅属于一个伪文档。假设有 K 个主题 $\{\phi_z\}_{z=1}^{K}$，每一个主题是在 V 维词典上的多项分布。为了能够自动判断词是 unigram 还是属于一个 collocaiton，我们进一步引入了 $K \times V$

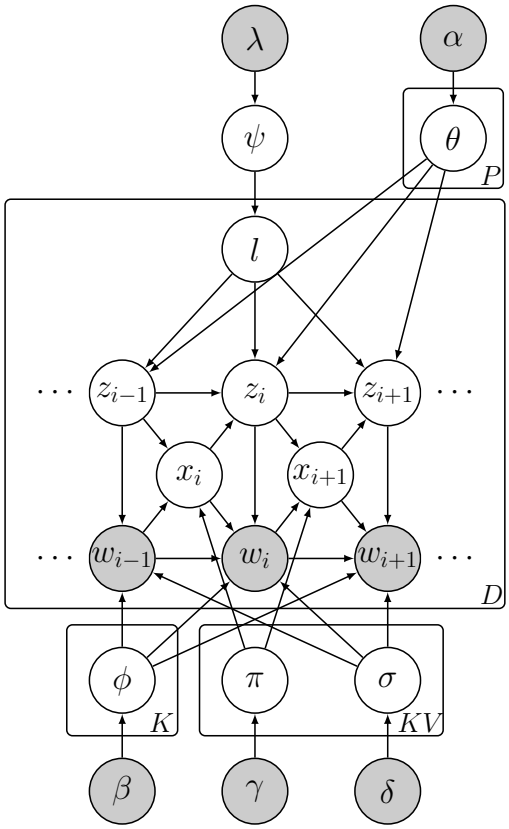


图 17　PTNG 的盘子表示法

个 Bernoulli 分布 $\pi_{z,w'}$ 以及 $K \times V$ 个在 V-维词典上的多项分布 $\sigma_{z,w'}$。为了生成短文本 d_s 中的词 w（非首个词），需要先从 $\pi_{z',w'}$ 采样一个 bigram 状态 x，其中 z' 是 w' 分配到的主题而 w' 是词 w 前面的一个词。至于短文本的首个词，我们固定它的 bigram 状态等于 0。如果一个词的 bigram 状态等于 0，我们从 θ_l 采样它的主题，然后从 ϕ_z 中采样该词。如果一个词的 bigram 状态等于 1，那么我们约束它的主题等于 w' 的主题，并且从 $\sigma_{z,w'}$ 中采样该词。换言之，x 等于 0 意味着 w 是一个 unigram，它的生成过程和 LDA 的一致。否则，它和它的前一个词构成 bigram。值得注意的是，连续的 bigram 可以形成 n-gram。因此 PTNG 可以自动检测短语或 collocations。

图 17给出了 PTNG 的盘子表示法。表 10给出了 PTNG 相关的数学符号和对应的解释。PTNG 的完整生成过程如下：

1. 采样 $\psi \sim \mathrm{Dir}(\lambda)$
2. 为每一个伪文档 d'_l:
    a. 采样 $\theta_l \sim \mathrm{Dir}(\alpha)$
3. 为每一个主题 z:
    a. 采样 $\phi_z \sim \mathrm{Dir}(\beta)$
4. 为每一个 z 以及每一个词 w:

表 10　PTNG 模型的数学符号表

符号	描述
K	主题数量
P	伪文档数量
D	短文本数量
V	词典大小
l	分配给一篇短文本的伪文档编号
z	分配给一个词的主题编号
x	两个连续词之间的 bigram 状态
w	一个观测到的词
ϕ_z	在词上的多项分布 w.r.t 主题 z
θ_l	在主题上的多项分布 w.r.t 伪文档 d'_l
ψ	伪文档在短文本上的多项分布
$\pi_{z,w}$	bigram 状态的 Bernoulli 分布 w.r.t. 前一个主题 z 和前一个词 w
$\sigma_{z,w}$	词上的多项分布 w.r.t. 前一个主题 z 和前一个词 w
λ	ψ 的 Dirichlet 先验参数
α	θ 的 Dirichlet 先验参数
β	ϕ 的 Dirichlet 先验参数
γ	π 的 Beta 先验参数
δ	σ 的 Dirichlet 先验参数
Bern$(\cdot)$	Bernoulli 分布
Beta$(\cdot)$	Beta 分布
Multi$(\cdot)$	Multinomial 分布
Dir$(\cdot)$	Dirichlet 分布

a. 采样 $\pi_{z,w} \sim \text{Beta}(\gamma)$

b. 采样 $\sigma_{z,w} \sim \text{Dir}(\delta)$

5. 为每一个短文本 d_s:

a. 采样一个伪文档 $l \sim \text{Multi}(\psi)$

b. 为 d_s 中的每一个词 w_i:

i. 采样 $x_i \sim \text{Bern}(\pi_{z_{i-1},w_{i-1}})$

ii. 如果 $x_i = 0$, 则采样 $z_i \sim \text{Multi}(\theta_l)$,
否则 $z_i = z_{i-1}$

iii. 如果 $x_i = 0$, 则采样 $w_i \sim \text{Multi}(\phi_{z_i})$,
否则采样 $w_i \sim \text{Multi}(\sigma_{z_{i-1},w_{i-1}})$

5.3.2 模型讨论

下面我们从 PTNG 和 TNG 之间的区别、为什么不基于其他短文本主题模型引入词序信息以及 PTNG 的模型复杂度三个方面对 PTNG 进行讨论：

1. 尽管 PTNG 发现 collocation 的机制和 Wang 等人 [99] 提出的 TNG 类似，但是二者之间有一个显著的区别。TNG 并不约束 collocation 中的词属于同一个主题，即图 17中没有 $z_{i-1} \rightarrow z_i$ 的边以及 $x_i \rightarrow z_i$ 的边。相对的，PTNG 约束 collocation 中的词属于同一个主题。在我们的实践中，去掉 $z_{i-1} \rightarrow z_i$ 以及 $x_i \rightarrow z_i$ 会严重降低 PTNG 的性能。

2. 基于 collocation 引入词序信息的主题模型面临更为严重的内容稀疏问题，因为相邻的词被拼接为 collocaiton 减少了共现信息。因此，在设计对词序敏感的主题模型之前，需要首先解决内容稀疏问题。直接建模词共现的主题模型 [23, 104, 105] 虽然可以解决内容稀疏问题，但是并不适用于本文的任务。因为这些方法不再建模短文本本身的生成过程，所以词序信息在构建 biterm 或者词网络的过程中已经损失了。尽管带有稀疏先验的主题模型 [24] 也可以引入 collocaiton 识别机制，但是稀疏模型本身的性能并不好。因此，我们最终采用伪文档主题模型作为基础模型引入词序信息。

3. 回顾一下 PTNG 中生成词的过程。当 bigram 状态 $x_i = 1$ 时，词 w_i 的生成是由一个条件概率分布 $p(w_i|z_{i-1}, w_{i-1})$ 定义。该分布有 $K \times V \times V$ 个参数，记为 $\sigma_{z_{i-1},w_{i-1}}$。可能有人会认为 $\sigma_{z,w}$ 太大以至无法放进内存。但是，实际的文本数据中，一个词只可能和少部分词出现在同一个上下文中。因此，$\sigma_{z,w}$ 是极其稀疏的。在实际的实现中，我们采用哈希表存储它。因此，引入 bigram 状态后，实际的 PTNG 模型复杂度相比于伪文档主题模型的复杂度并没有提高太多。

5.3.3 统计推断

对于 PTNG 而言，精确的后验推断是不可行。所以我们采用 Gibbs 采样进行近似后验推断。由于采用了共轭先验，我们可以轻松积分掉 θ, ϕ, ψ, π 以及 σ。这有助于 Gibbs 采样的稳定和快速收敛。需要采样的隐变量为伪文档分配 l，主题分配 z 以及 bigram 状态 x。下面我们给出这些隐变量的采样公式。

采样伪文档 l。给定其他变量的值后，采样 l 和 DMM 模型 [96] 中采样主题的方法类

似：

$$\begin{aligned} p(l_{d_s} = l|rest) &\propto \frac{M_{l,\neg d_s}}{D-1+P\lambda} \frac{\frac{\prod_{z\in d_s}\Gamma\left(N_l^z+\alpha\right)}{\prod_{z\in d_s}\Gamma\left(N_{l,\neg d_s}^z+\alpha\right)}}{\prod_{i=1}^{N_{d_s}}\left(N_{l,\neg d_s}+K\alpha+i-1\right)} \\ &= \frac{M_{l,\neg d_s}}{D-1+P\lambda} \frac{\prod_{z\in d_s}\prod_{j=1}^{N_{d_s}^z}\left(N_{l,\neg d_s}^z+\alpha+j-1\right)}{\prod_{i=1}^{N_{d_s}}\left(N_{l,\neg d_s}+K\alpha+i-1\right)}, \end{aligned} \tag{5.1}$$

其中 M_l 是分配到第 l 个伪文档 d_l' 中的短文本数量。N_{d_s} 是第 s 个短文本 d_s 中词的数量，$N_{d_s}^z$ 是 d_s 中被分配到主题 z 下词的数量。类似的，N_l 是伪文档 d_l' 中词的数量，N_l^z 是 d_l' 中分配到主题 z 下词的数量。所有的带有下标 $\neg d_s$ 的计数都意味着没有计入 d_s 中相应计数的结果。

采样主题 z 以及 bigram 状态 x。 我们将一个词对应的主题 z 和 bigram 状态 x 当作一个整体进行采样，而不是分别采样：

$$\begin{aligned} &p(z_{d_s,i} = z_i, x_{d_s,i} = x_i|rest) \\ &\propto (\gamma + N_{z_{i-1},w_{i-1}}^{x_i} - 1)(\alpha + N_{l_{d_s}}^{z_i} - 1) \\ &\times \begin{cases} \frac{N_{z_i}^{w_i}+\beta-1}{N_{z_i}+V\beta-1} & \text{if } x_i = 0 \\ \frac{N_{z_i,w_{i-1}}^{w_i}+\delta-1}{N_{z_i,w_{i-1}}+V\delta-1} & \text{if } x_i = 1 \& z_i = z_{i-1} \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{5.2}$$

其中 N_z^w 是词被分配到主题 z 中的次数，$N_{z,w'}^w$ 是词 w 和前一个词 w' 形成 bigram 并且被分配到主题 z 中的次数。$N_z = \sum_{w=0}^V N_z^w$，$N_{z,w'} = \sum_{w=0}^V N_{z,w'}^w$。$N_{z_{i-1},w_{i-1}}^{x_i}$ 是当 $x_i = 1$ (或 $x_i = 0$) 时，词 w_{i-1} 和任何其他词形成 bigram 并且被分配到主题 z 中的次数。

为了方便讨论，假设每篇短文本包含 N 个词。在 PTNG 中，为每个短文本采样伪文档 l 的时间复杂度为 $O(P)$，采样 (z,x) 的时间复杂度为 $O(2NK)$。那么，PTNG 的 Gibbs 采样算法在一篇短文本上总的时间复杂度为 $O(P+2NK)$。因此，PTNG 可以有效地应用到数万深知数十万的短文本数据上。

有了采样收敛后得到的 l、z 和 x，很容易通过经验估计得到 θ, ϕ, ψ, π 以及 σ 的值。下面以 θ 和 ϕ 为例，给出估计公式：

$$\theta_{l,z} = \frac{N_l^z + \alpha}{N_l + K\alpha} \tag{5.3}$$

$$\phi_{z,w} = \frac{N_z^w + \beta}{N_z + V\beta} \tag{5.4}$$

值得注意的是，PTNG 学习到的 θ_l 是伪文档 d_l' 的主题分布。对于短文本 d_s 的主题

表 11　数据集的统计指标

Data set	# Documents	Vocabulary size	Avg. document length
News	29,200	11,007	12.4
DBLP	55,290	7,525	6.4
Tweets	182,671	21,480	8.5

分布 θ_s，我们需要利用 Gibbs 采样得到的 $\vec{z}$，通过经验估计得到：

$$\theta_{s,z} = \frac{N_{d_s}^{z} + \alpha}{N_{d_s} + K\alpha}, \tag{5.5}$$

其中 $N_{d_s}^{z}$ 是短文本 d_s 中被分配给主题 z 的词数。

对于某些应用，主题模型的性能和超参数的设置有关。在本文讨论的实验中，我们发现超参数对结果的影响有限。因此，我们并不实际学习超参数的值，而是直接为它们设置经验值。

5.4　实验结果与分析

5.4.1　实验设置

数据集

为了测试 PTNG 的性能，我们在三个真实的短文本数据集进行主题一致性以及短文本分类的实验。表 11给出了它们的一些统计结果。下面分别对它们作简单介绍。

News： 该数据集[1]包含 29200 条英文新闻。每条新闻属于一个类别。所有类别包括 sport, business, U.S., health, sci&tech, world 和 entertainment。我们仅使用新闻的摘要，因为它是典型的短文本。

DBLP： 我们收集了计算机学科 6 个研究领域会议论文的标题。具体的领域包括：数据挖掘、机器视觉、数据库、信息检索、自然语言处理和机器学习。最终获得了 55290 个短文本，每个都标注了 6 个研究领域中的一个。

Tweets： Zubiaga 等人 [97] 抓取并标注了大量的 tweets。他们抓取了包含 URL 的 tweets，并且用 URL 指向的网页的类别作为对应 tweet 的类别。网页的类别是根据开放目录计划（Open Directory Project 或 ODP）定义的。该数据集包含 10 个不同类别，一共 360k 左右的 tweets。我们从中选择了 9 个话题相关的类别，并从中抽样了 182671 条

[1] http://acube.di.unipi.it/tmn-dataset/

tweets。

基准方法

Latent Dirichlet Allocation (LDA)。作为最经典的主题模型，LDA [15] 可以通过将 Dirichlet 参数设置的趋近于零来保证模型的稀疏性。这里 LDA 的实现使用的是开源的 jGibbLDA[2]。

Mixture of Unigrams (MU)。MU [22] 最大的特点是假设每个文档所有的词都只属于一个主题。MU 通过该假设强制每篇文档的主题表示达到最大的稀疏度。对于长文本，该假设也许不太合理，但是在短文本上也许是可行的。

Dual Sparse Topic Model (DSTM)。DSTM [24] 是近期提出的带有稀疏先验的主题模型。它将 LDA 的文档-主题分布以及主题-词分布的 Dirichlet 先验均修改为 “Spike-and-Slab” 稀疏先验。这使得 DSTM 可以适用于短文本一类的稀疏数据。

Topical N-gram model (TNG)。 - 本实验中使用的 TNG [99] 和原始论文中有一点区别。因为我们对其进行了修改，约束短语中的每个词共享一个主题。这个修改极大地提高了 TNG 在短文本上的性能。

Pseudo-document-based Topic Model (PTM)。- PTM [106] 就是第 四章中介绍的伪文档主题模型。通过将短文本聚合为伪文档，PTM 在短文本主题建模上取得了不错的效果。但是，它并没有引入词序信息。

参数设置

我们设置 LDA 和 MU 的 $\alpha = 0.1$，$\alpha = 0.1$，因为带有弱先验的它们在短文本上表现更好。我们设置 DSTM 的 $\pi = 0.1$，$\gamma = 0.01$，发现比原论文中的 $\pi = 1.0$ 以及 $\gamma = 1.0$ 表现更好。至于 DSTM 中的 $\bar{\pi}$ 和 $\bar{\gamma}$，我们采用原论文的设置。对于 PTM、PTNG 和 TNG，我们设置 $\alpha = 0.1$，$\beta = 0.01$。对于 PTNG 和 PTM，我们设置 $\lambda = 0.1$。对于 PTNG 以及 TNG，我们设置 $\gamma = 0.1$，$\delta = 0.01$。

对于所有的主题模型，我们统一设置主题数 $K = 100$。PTNG 和 PTM 的伪文案数统一设置为 1000。除了 PTNG 和 TNG 的采样迭代了 5000 轮，其他方法均迭代 2000 次。所有实验结果均是五次实验结果的均值。

评价指标

主题一致性（Topic Coherence）。我们使用 UCI 一致性指标评价 News 和 DBLP 的

[2]http://jgibblda.sourceforge.net

结果。UCI 一致性指标使用点互信息（PMI）计算主题的一致性。给定一个主题 z，我们选择它的 top-N 主题词 $w_1, w_2, \cdots, w_N$，然后计算这些词构成的每个词对的 PMI 的均值：

$$PMI(z) = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{5.6}$$

其中 $p(w_i, w_j)$ 是词对 (w_i, w_j) 出现在同一个滑动窗口的联合概率，$p(w_i)$ 是词 w_i 出现在一个滑动窗口的边缘概率。这些概率值是在 Wikipedia 的文档上统计得到的。我们使用所有主题的一致性指标的均值作为一个主题模型的主题一致性结果。实验中 N 的值统一设置为 10。

分类指标。主题模型一个常用的外部评价方式是文档分类。因此，我们使用主题模型学到的短文本语义表示去训练分类器，然后使用宏平均的精确度、召回率和 F-值评价分类效果。

5.4.2 实验结果

UCI 主题一致性结果

图 18给出了 PTNG 和所有基准方法在 News 和 DBLP 上的 UCI 主题一致性结果。由于 UCI 主题一致性比 UMass 主题一致性更适用于短文本，但是它需要在合适的外部语料上统计词共现概率。对于 News 和 DBLP 这类经过仔细编辑的文本，我们可以放心地使用维基百科作为外部语料。但是，对于用户产生的 tweets，很难找到合适的外部语料。因此，我们只在 News 和 DBLP 上进行了主题一致性的对比。

从图 18中不难发现，PTNG 在 News 和 DBLP 两个数据集上的主题一致性均超过了所有基准方法。PTNG 比 PTM 更能学到高质量的主题，这说明了将词序引入短文本主题建模过程中是有用的。类似的，PTNG 的效果超过 TNG 说明了短文本聚合对解决短文本上 N-gram 主题建模时遇到的内容稀疏问题是有效的。在 News 上 TNG 的效果和 LDA 类似，而在 DBLP 上略微超过 LDA。这表明了，仅仅引入词序信息对短文本主题建模效果难有提升。因此必须同时引入词序信息和进行短文本聚合。一个有趣的现象是带有弱先验的 LDA 比 DSTM 的效果好。后者为文档-主题分布以及主题-词分布分别使用了稀疏先验，被认为理论上比 LDA 更适用于短文本。不过，稀疏先验中引入的额外分布可能导致了模型在实际短文本上训练比较困难。

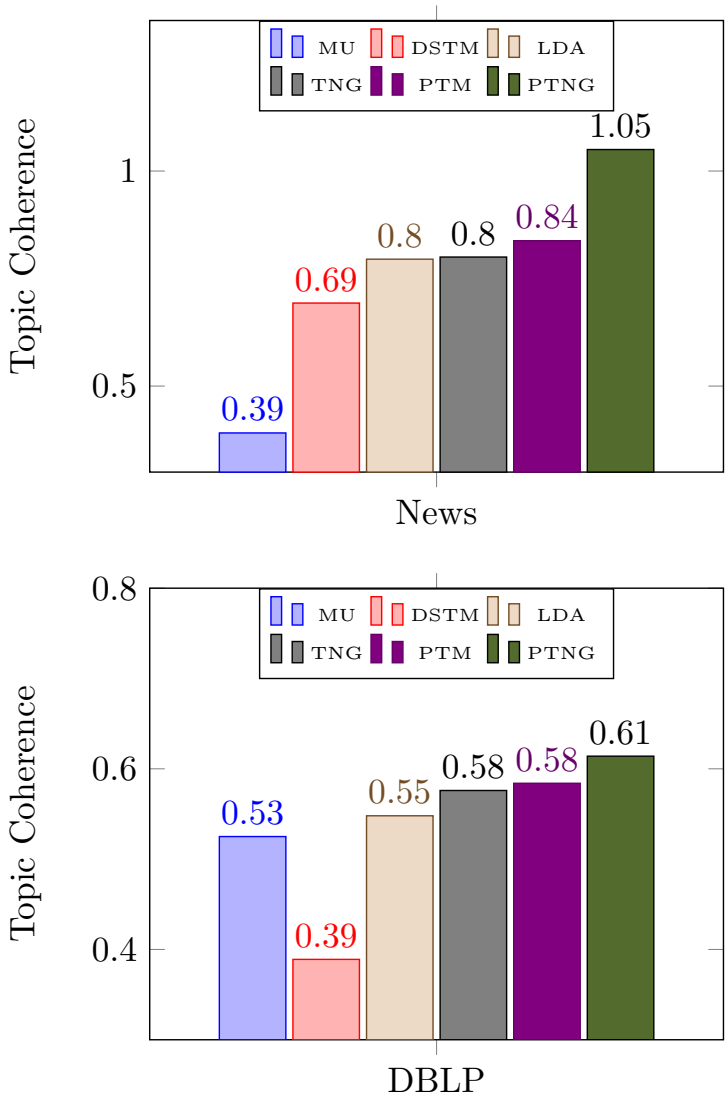


图 18　News 以及 DBLP 上 UCI 主题一致性结果

表 12　五折交叉分类实验结果

	News			DBLP			Tweets		
	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
PTNG	**0.770**	**0.770**	**0.769**	*0.662*	*0.669*	*0.664*	*0.597*	**0.604**	**0.596**
PTM	*0.755*	*0.757*	*0.754*	**0.667**	**0.672**	**0.668**	0.561	0.568	*0.559*
TNG	0.710	0.715	0.710	0.608	0.619	0.610	0.559	*0.571*	0.558
LDA	0.727	0.732	0.728	0.613	0.624	0.614	0.553	0.560	0.546
DSTM	0.720	0.724	0.720	0.619	0.628	0.620	0.539	0.547	0.535
MU	0.697	0.617	0.626	0.640	0.643	0.638	**0.634**	0.546	0.546

5.4.3 短文本分类结果

我们对比所有主题主题模型在短文本分类上的效果。这里，我们将主题模型用作一种降维工具，并将短文本表示为主题分布。用主题分布作为特征，训练分类器。每个主题模型训练得到所有短文本的主题表示后进行五折交叉验证。分类器采用 LIBLINEAR[3]。表 12给出了宏平均的准确度、召回率以及 F-值。最好的结果用黑体表示，次好的结果用斜体表示。

从结果中不难发现，PTM 在 News、DBLP 以及 Question 数据上均是所有基准方法中表现最好的。这说明了短文本聚合为长文本可以有效地提高短文本主题建模的效果。我们提出的 PTNG 在 News 以及 Tweets 上取得了比 PTM 更好的效果，而在 DBLP 上和 PTM 的效果很接近。PTNG 良好的表现说明了词序在准确地学习短文本主题中的关键作用，这进一步保证了短文本主题表示的可区分度。TNG 的表现在 News 和 DBLP 上略逊于 LDA，这表明了只考虑词序信息并不能解决短文本主题建模的问题。MU 和 DSTM 的表现不佳说明了加稀疏先验在短文本上的实际效果有限。

5.5 小结

本节我们提出了伪文档 N-gram 主题模型（ PTNG ）。通过自动发现 collocaton, PTNG 将词序信息引入短文本主题建模。此外，通过短文本自聚合，PTNG 成功缓解了短文本内容稀疏问题。据我们所知，PTNG 是首个尝试在短文本主题建模中考虑词序的工作。在三个真实短文本数据集上得到的实验结果表明，PTNG 在主题一致性和短文本分类任务上均取得了很好的效果。实验结果还表明单独引入词序信息并不能提高短文本主题建模的效果，必须结合短文本自聚合方法来解决短文本内容稀疏问题。

[3]http://www.csie.ntu.edu.tw/ cjlin/liblinear/

第六章　总结与展望

6.1　本文工作总结

自经典概率主题模型流行以来已有十多年时间，基于它的扩展模型以及各类应用层出不穷。但是，绝大多数工作都是基于长文本。随着互联网上用户产生的短文本越来越多，有研究者发现现有主题模型直接应用到短文本时效果一般，不如先将短文本聚合为长文本后训练得到的主题模型效果好。近年来，针对短文本设计的主题模型开始出现。但是，现有的模型存在各自的局限性。此外，现有模型都仅仅着力于解决短文本内容稀疏的问题，忽略了词序对短文本主题建模的帮助。

因此，本文开展了针对短文本的主题建模研究，以提高短文本语义分析与处理的能力。针对现有模型的局限性，以及考虑到词序对确定短文本语义的帮助作用，我们展开了三方面的研究。

- 经典主题模型建模的是文档的生成过程，它假设每篇文档都对应了一个主题分布。文档中的词共现信息是学习该分布的主要依据。但是，短文本上词很稀疏，导致词的文档级共现信息十分缺乏。解决上述问题的一个直接方法是不再建模文档的生成过程，改为直接建模文档集合级别的词共现信息。这类方法的代表是双词主题模型。由于双词主题模型过于简单，使得它的实际效果有限。采用直接建模词共现的思路，我们设计了表达能力更强的词网络主题模型。具体的，我们首先将短文本转换为词共现网络，然后建模词共现网络的生成过程。通过这种方式学习主题避免了生成短文本时面临的内容稀疏问题。
- 将短文本转换为词共现网络增加了数据的稠密度，有助于学习短文本中的主题。但是，由于不再生成原始短文本，所以失去了直接建模短文本主题分布的能力。因此，我们提出伪文档主题模型。通过短文本聚合，它可以解决短文本内容稀疏问题，同时学习短文本本身的主题分布。短文本聚合可以创造跨原始短文本的词共现信息，因此在一定程度上能够解决内容稀疏问题。此外，短文本本身的生成过程是包含在伪文档的生成过程中的。因此，伪文档主题模型能够在建模短文本自身主题分布的同时，更准确地学习主题。
- 现有短文本主题模型仍然采用词袋模型的假设，完全忽略了词序信息。但是，词序信息对于准确地学习短文本中的主题是有帮助的。目前，将词序信息引入主题模型的主要方式是同时学习主题和 collocaitons。现有基于 collocation 的主题模型

直接应用到短文本会加剧内容稀疏问题。因此，将词序引入短文本主题建模需要设计特殊的模型。我们提出了伪文档 N-gram 模型，一方面通过短文本自聚合解决短文本内容稀疏问题，另一方面通过学习 collocation 引入词序信息。

6.2　未来工作展望

在我们的研究过程中，为了保证新方法可以应用到所有类型的短文本上，忽略了特定类型短文本可能拥有的特征或者上下文信息。此外，随着技术的发展和进步，越来越多新的方法可以用来帮助提高短文本主题建模的能力。

- 结合上下文信息 - 真实的短文本除了文本信息以外，还有很多非文本的上下文信息。例如，作者/用户、时间、地点和人物关系等。结合这些上下文，可以更全面地描述文本内容从而更准确地刻画主题。
- 结合词向量方法 - 词向量方法在文本处理中得到越来越多的应用。它可以将词表示为潜在空间中的点，相近的点拥有相近的语义。词向量应用于短文本的最大优势是可以在大规模语料上预训练词向量。我们将考虑利用在大规模外部语料上训练得到的词向量去缓解短文本的内容稀疏问题。
- 结合深度学习方法 - 近几年深度学习在自然语言处理领域应用得越来越多。其中循环神经网络由于出色的序列建模能力被广泛用于语言模型。我们将考虑结合循环神经网络设计短文本主题模型，来提高对短文本词序信息的建模能力。

参 考 文 献

[1] STEYVERS M, GRIFFITHS T. Probabilistic Topic Models[M]LANDAUER T, MCNAMARA D, DENNIS S, et al. Handbook of Latent Semantic Analysis : Lawrence Erlbaum Associates, 2007. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1410615340.

[2] ARAMPATZIS A, KAMPS J. A Study of Query Length[C]. SIGIR '08 : Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 2008 : 811 – 812. http://doi.acm.org/10.1145/1390334.1390517.

[3] SALTON G, WONG A, YANG C S. A Vector Space Model for Automatic Indexing[J]. Commun. ACM, 1975, 18(11) : 613 – 620. http://doi.acm.org/10.1145/361219.361220.

[4] PONTE J M, CROFT W B. A Language Modeling Approach to Information Retrieval[C]. SIGIR '98 : Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 1998 : 275 – 281. http://doi.acm.org/10.1145/290941.291008.

[5] METZLER D, DUMAIS S, MEEK C. Similarity Measures for Short Segments of Text[C]. ECIR'07 : Proceedings of the 29th European Conference on IR Research. Berlin, Heidelberg : Springer-Verlag, 2007 : 16 – 27. http://dl.acm.org/citation.cfm?id=1763653.1763660.

[6] XU Y, JONES G J, WANG B. Query Dependent Pseudo-relevance Feedback Based on Wikipedia[C]. SIGIR '09 : Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 2009 : 59 – 66. http://doi.acm.org/10.1145/1571941.1571954.

[7] SAHAMI M, HEILMAN T D. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets[C]. WWW '06 : Proceedings of the 15th International Conference on World Wide Web. New York, NY, USA : ACM, 2006 : 377 – 386. http://doi.acm.org/10.1145/1135777.1135834.

[8] BANERJEE S, RAMANATHAN K, GUPTA A. Clustering Short Texts Using

Wikipedia[C]. SIGIR '07 : Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 2007 : 787 – 788. http://doi.acm.org/10.1145/1277741.1277909.

[9] HU X, SUN N, ZHANG C, et al. Exploiting Internal and External Semantics for the Clustering of Short Texts Using World Knowledge[C]. CIKM '09 : Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York, NY, USA : ACM, 2009 : 919 – 928. http://doi.acm.org/10.1145/1645953.1646071.

[10] SONG Y, WANG H, WANG Z, et al. Short Text Conceptualization Using a Probabilistic Knowledgebase[C]. IJCAI'11 : Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three : AAAI Press, 2011 : 2330 – 2336. http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-388.

[11] YAN X, GUO J, LIU S, et al. Clustering Short Text Using Ncut-weighted Non-negative Matrix Factorization[C]. CIKM '12 : Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, NY, USA : ACM, 2012 : 2259 – 2262. http://doi.acm.org/10.1145/2396761.2398615.

[12] PHAN X-H, NGUYEN L-M, HORIGUCHI S. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections[C]. WWW '08 : Proceedings of the 17th International Conference on World Wide Web. New York, NY, USA : ACM, 2008 : 91 – 100. http://doi.acm.org/10.1145/1367497.1367510.

[13] CHEN M, JIN X, SHEN D. Short Text Classification Improved by Learning Multi-granularity Topics[C]. IJCAI'11 : Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three : AAAI Press, 2011 : 1776 – 1781. http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-298.

[14] YU H-F, HO C-H, JUAN Y-C, et al. LibShortText: A Library for Short-text Classification and Analysis LibShortText: A Library for Short-text Classification and Analysis[R]. 2013.

[15] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. J. Mach. Learn. Res., 2003, 3 : 993 – 1022. http://dl.acm.org/citation.cfm?id=944919.944937.

[16] TANG J, MENG Z, NGUYEN X, et al. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis[C]. ICML'14 : Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 : JMLR.org, 2014 : I – 198. http://dl.acm.org/citation.cfm?id=3044805.3044828.

[17] HONG L, DAVISON B D. Empirical Study of Topic Modeling in Twitter[C]. SOMA '10 : Proceedings of the First Workshop on Social Media Analytics. New York, NY, USA : ACM, 2010 : 80 – 88. http://doi.acm.org/10.1145/1964858.1964870.

[18] MEHROTRA R, SANNER S, BUNTINE W, et al. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling[C]. SIGIR '13 : Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 2013 : 889 – 892. http://doi.acm.org/10.1145/2484028.2484166.

[19] QUAN X, KIT C, GE Y, et al. Short and Sparse Text Topic Modeling via Self-aggregation[C]. IJCAI'15 : Proceedings of the 24th International Conference on Artificial Intelligence : AAAI Press, 2015 : 2270 – 2276. http://dl.acm.org/citation.cfm?id=2832415.2832564.

[20] TANG J, ZHANG M, MEI Q. One Theme in All Views: Modeling Consensus Topics in Multiple Contexts[C]. KDD '13 : Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA : ACM, 2013 : 5 – 13. http://doi.acm.org/10.1145/2487575.2487682.

[21] JIN O, LIU N N, ZHAO K, et al. Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering[C]. CIKM '11 : Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York, NY, USA : ACM, 2011 : 775 – 784. http://doi.acm.org/10.1145/2063576.2063689.

[22] NIGAM K, MCCALLUM A K, THRUN S, et al. Text Classification from Labeled and Unlabeled Documents Using EM[J]. Mach. Learn., 2000, 39(2-3) : 103 – 134. http://dx.doi.org/10.1023/A:1007692713085.

[23] YAN X, GUO J, LAN Y, et al. A Biterm Topic Model for Short Texts[C]. WWW '13 : Proceedings of the 22Nd International Conference on World Wide Web. New York, NY, USA : ACM, 2013 : 1445 – 1456. http://doi.acm.org/10.1145/2488388.2488514.

[24] LIN T, TIAN W, MEI Q, et al. The Dual-sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text[C]. WWW '14 : Proceedings of the 23rd International Conference on World Wide Web. New York, NY, USA : ACM, 2014 : 539 – 550. http://doi.acm.org/10.1145/2566486.2567980.

[25] WALLACH H M. Topic Modeling: Beyond Bag-of-words[C]. ICML '06 : Proceedings of the 23rd International Conference on Machine Learning. 2006 : 977 – 984.

[26] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 1990, 41(6) : 391 – 407.

[27] HOFMANN T. Probabilistic Latent Semantic Indexing[C]. SIGIR. 1999 : 50 – 57.

[28] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 1977, 39(1) : 1 – 38.

[29] LEE D D, SEUNG H S. Learning the parts of objects by nonnegative matrix factorization[J]. Nature, 1999, 401 : 788 – 791.

[30] GAUSSIER E, GOUTTE C. Relation Between PLSA and NMF and Implications[C]. SIGIR '05 : Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 2005 : 601 – 602. http://doi.acm.org/10.1145/1076034.1076148.

[31] LI T, DING C. The Relationships Among Various Nonnegative Matrix Factorization Methods for Clustering[C]. ICDM '06 : Proceedings of the Sixth International Conference on Data Mining. Washington, DC, USA : IEEE Computer Society, 2006 : 362 – 371. http://dx.doi.org/10.1109/ICDM.2006.160.

[32] CAI D, HE X, WU X, et al. Non-negative Matrix Factorization on Manifold[C]. Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy. 2008 : 63 – 72. http://dx.doi.org/10.1109/ICDM.2008.57.

[33] ZHU J, XING E P. Sparse Topical Coding[C]. UAI'11 : Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia, United States : AUAI Press, 2011 : 831 – 838. http://dl.acm.org/citation.cfm?id=3020548.3020644.

[34] WANG Q, CAO Z, XU J, et al. Group Matrix Factorization for Scalable Topic Modeling[C]. SIGIR '12 : Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 2012 : 375 – 384. http://doi.acm.org/10.1145/2348283.2348335.

[35] WANG Q, XU J, LI H, et al. Regularized Latent Semantic Indexing[C]. SIGIR '11 : Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 2011 : 685 – 694. http://doi.acm.org/10.1145/2009916.2010008.

[36] CHEN X, QI Y, BAI B, et al. Sparse Latent Semantic Analysis.[C]. SDM : SIAM / Omnipress, 2011 : 474 – 485. http://dblp.uni-trier.de/db/conf/sdm/sdm2011.html#ChenQBLC11.

[37] LAROCHELLE H, LAULY S. A Neural Autoregressive Topic Model[G]PEREIRA F, BURGES C J C, BOTTOU L, et al. Advances in Neural Information Processing Systems 25 : Curran Associates, Inc., 2012 : 2708 – 2716. http://papers.nips.cc/paper/4613-a-neural-autoregressive-topic-model.pdf.

[38] MAAS A L, NG A Y. A probabilistic model for semantic word vectors[C]. Workshop on Deep Learning and Unsupervised Feature Learning, NIPS : Vol 10. 2010.

[39] HINTON G E, SALAKHUTDINOV R R. Replicated Softmax: an Undirected Topic Model[G]BENGIO Y, SCHUURMANS D, LAFFERTY J D, et al. Advances in Neural Information Processing Systems 22 : Curran Associates, Inc., 2009 : 1607 – 1614. http://papers.nips.cc/paper/3856-replicated-softmax-an-undirected-topic-model.pdf.

[40] SARIKAYA R, HINTON G E, DEORAS A. Application of Deep Belief Networks for Natural Language Understanding[J]. IEEE/ACM Trans. Audio, Speech and Lang. Proc., 2014, 22(4) : 778 – 784. http://dx.doi.org/10.1109/TASLP.2014.2303296.

[41] SOCHER R, BENGIO Y, MANNING C D. Deep Learning for NLP (Without Magic)[C]. ACL '12 : Tutorial Abstracts of ACL 2012. Stroudsburg, PA, USA : Association for Computational Linguistics, 2012 : 5 – 5. http://dl.acm.org/citation.cfm?id=2390500.2390505.

[42] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences, 2004, 101(Suppl. 1) : 5228 – 5235.

[43] MINKA T, LAFFERTY J. Expectation-propagation for the Generative Aspect Model[C]. UAI'02 : Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2002 : 352 – 359. http://dl.acm.org/citation.cfm?id=2073876.2073918.

[44] CHIEN J-T, WU M-S. Adaptive Bayesian Latent Semantic Analysis[J]. Trans. Audio, Speech and Lang. Proc., 2008, 16(1) : 198 – 207. http://dx.doi.org/10.1109/TASL.2007.909452.

[45] ASUNCION A, WELLING M, SMYTH P, et al. On Smoothing and Inference for Topic Models[C]. UAI '09 : Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia, United States : AUAI Press, 2009 : 27 – 34.

http://dl.acm.org/citation.cfm?id=1795114.1795118.

[46] GEMAN S, GEMAN D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images[J]. IEEE Trans. Pattern Anal. Mach. Intell., 1984, 6(6): 721–741. http://dx.doi.org/10.1109/TPAMI.1984.4767596.

[47] HEINRICH G. Parameter estimation for text analysis[R]. http://www.arbylon.net/publications/text-est.pdf: vsonix GmbH and University of Leipzig, 2008.

[48] YAO L, MIMNO D, MCCALLUM A. Efficient Methods for Topic Model Inference on Streaming Document Collections[C]. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009: 937–946.

[49] LI A Q, AHMED A, RAVI S, et al. Reducing the Sampling Complexity of Topic Models[C]. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 891–900.

[50] WANG Y, BAI H, STANTON M, et al. PLDA: Parallel Latent Dirichlet Allocation for Large-Scale Applications[C]. AAIM '09: Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management. Berlin, Heidelberg: Springer-Verlag, 2009: 301–314. http://dx.doi.org/10.1007/978-3-642-02158-9_26.

[51] LIU Z, ZHANG Y, CHANG E Y, et al. PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing[J]. ACM Trans. Intell. Syst. Technol., 2011, 2(3): 26:1–26:18. http://doi.acm.org/10.1145/1961189.1961198.

[52] SMYTH P, WELLING M, ASUNCION A U. Asynchronous Distributed Learning of Topic Models[G]KOLLER D, SCHUURMANS D, BENGIO Y, et al. Advances in Neural Information Processing Systems 21: Curran Associates, Inc., 2009: 81–88. http://papers.nips.cc/paper/3524-asynchronous-distributed-learning-of-topic-models.pdf.

[53] AZZOPARDI L, GIROLAMI M, van RISJBERGEN K. Investigating the Relationship Between Language Model Perplexity and IR Precision-recall Measures[C]. SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. New York, NY, USA: ACM, 2003: 369–370. http://doi.acm.org/10.1145/860435.860505.

[54] CHANG J, GERRISH S, WANG C, et al. Reading Tea Leaves: How Humans Interpret Topic Models[C]. NIPS. 2009: 288–296.

[55] BLEI D M. Probabilistic Topic Models[J]. Commun. ACM, 2012, 55(4): 77–84. http://doi.acm.org/10.1145/2133806.2133826.

[56] ZHAO W X, JIANG J, WENG J, et al. Comparing Twitter and Traditional Media Using Topic Models[C]. ECIR'11: Proceedings of the 33rd European Conference on Advances in Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 2011: 338–349. http://dl.acm.org/citation.cfm?id=1996889.1996934.

[57] NEWMAN D, LAU J H, GRIESER K, et al. Automatic Evaluation of Topic Coherence[C]. HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010: 100–108. http://dl.acm.org/citation.cfm?id=1857999.1858011.

[58] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing Semantic Coherence in Topic Models[C]. EMNLP. 2011: 262–272.

[59] LU Y, MEI Q, ZHAI C. Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA[J]. Inf. Retr., 2011, 14(2): 178–203. http://dx.doi.org/10.1007/s10791-010-9141-9.

[60] CAI D, MEI Q, HAN J, et al. Modeling Hidden Topics on Document Manifold[C]. CIKM '08: Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York, NY, USA: ACM, 2008: 911–920. http://doi.acm.org/10.1145/1458082.1458202.

[61] BLEI D M, JORDAN M I, GRIFFITHS T L, et al. Hierarchical Topic Models and the Nested Chinese Restaurant Process[C]. NIPS'03: Proceedings of the 16th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2003: 17–24. http://dl.acm.org/citation.cfm?id=2981345.2981348.

[62] TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical Dirichlet Processes[J]. Journal of the American Statistical Association, 2006, 101(476): 1566–1581.

[63] WANG C, BLEI D M. Collaborative Topic Modeling for Recommending Scientific Articles[C]. KDD '11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2011: 448–456. http://doi.acm.org/10.1145/2020408.2020480.

[64] MIMNO D, WALLACH H M, NARADOWSKY J, et al. Polylingual Topic Models[C]. EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Lan-

guage Processing: Volume 2 - Volume 2. Stroudsburg, PA, USA : Association for Computational Linguistics, 2009 : 880 – 889. http://dl.acm.org/citation.cfm?id=1699571.1699627.

[65] NI X, SUN J-T, HU J, et al. Mining Multilingual Topics from Wikipedia[C]. WWW '09 : Proceedings of the 18th International Conference on World Wide Web. New York, NY, USA : ACM, 2009 : 1155 – 1156. http://doi.acm.org/10.1145/1526709.1526904.

[66] BLEI D M, MCAULIFFE J D. Supervised Topic Models[C]. NIPS. 2007 : 121 – 128.

[67] RAMAGE D, HALL D, NALLAPATI R, et al. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora[C]. EMNLP. 2009 : 248 – 256.

[68] LACOSTE-JULIEN S, SHA F, JORDAN M I. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification[G]KOLLER D, SCHUURMANS D, BENGIO Y, et al. Advances in Neural Information Processing Systems 21 : Curran Associates, Inc., 2009 : 897 – 904. http://papers.nips.cc/paper/3599-disclda-discriminative-learning-for-dimensionality-reduction-and-classification.pdf.

[69] ZHU J, AHMED A, XING E P. MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification[C]. ICML '09 : Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA : ACM, 2009 : 1257 – 1264. http://doi.acm.org/10.1145/1553374.1553535.

[70] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The Author-topic Model for Authors and Documents[C]. UAI. 2004 : 487 – 494.

[71] WANG X, MCCALLUM A. Topics over Time: A non-Markov Continuous-time Model of Topical Trends[C]. KDD '06 : Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA : ACM, 2006 : 424 – 433. http://doi.acm.org/10.1145/1150402.1150450.

[72] BOYD-GRABER J L, BLEI D M. Syntactic Topic Models[G]KOLLER D, SCHURMANS D, BENGIO Y, et al. Advances in Neural Information Processing Systems 21 : Curran Associates, Inc., 2009 : 185 – 192. http://papers.nips.cc/paper/3398-syntactic-topic-models.pdf.

[73] GRIFFITHS T L, STEYVERS M, BLEI D M, et al. Integrating Topics and Syntax[G]SAUL L K, WEISS Y, BOTTOU L. Advances in Neural Information Processing Systems 17 : MIT Press, 2005 : 537 – 544. http://papers.nips.cc/paper/

2587-integrating-topics-and-syntax.pdf.

[74] MEI Q, CAI D, ZHANG D, et al. Topic Modeling with Network Regularization[C]. WWW '08 : Proceedings of the 17th International Conference on World Wide Web. New York, NY, USA : ACM, 2008 : 101 – 110. http://doi.acm.org/10.1145/1367497.1367512.

[75] NALLAPATI R M, AHMED A, XING E P, et al. Joint Latent Topic Models for Text and Citations[C]. KDD '08 : Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA : ACM, 2008 : 542 – 550. http://doi.acm.org/10.1145/1401890.1401957.

[76] CHANG J, BLEI D M. Relational Topic Models for Document Networks.[C]DYK D A V, WELLING M. JMLR Proceedings, Vol 5 : AISTATS : JMLR.org, 2009 : 81 – 88.

[77] WENG J, LIM E-P, JIANG J, et al. TwitterRank: Finding Topic-sensitive Influential Twitterers[C]. WSDM '10 : Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York, NY, USA : ACM, 2010 : 261 – 270. http://doi.acm.org/10.1145/1718487.1718520.

[78] CHEN Y, AMIRI H, LI Z, et al. Emerging Topic Detection for Organizations from Microblogs[C]. SIGIR '13 : Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : ACM, 2013 : 43 – 52. http://doi.acm.org/10.1145/2484028.2484057.

[79] CHUA F C T, ASUR S. Automatic Summarization of Events from Social Media.[C]KICIMAN E, ELLISON N B, HOGAN B, et al. ICWSM : The AAAI Press, 2013.

[80] JAGARLAMUDI J, III DAUMÉ H, UDUPA R. Incorporating Lexical Priors into Topic Models[C]. EACL '12 : Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA : Association for Computational Linguistics, 2012 : 204 – 213. http://dl.acm.org/citation.cfm?id=2380816.2380844.

[81] ANDRZEJEWSKI D, ZHU X, CRAVEN M. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors[C]. ICML '09 : Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA : ACM, 2009 : 25 – 32. http://doi.acm.org/10.1145/1553374.1553378.

[82] WALLACH H M, MIMNO D M, MCCALLUM A. Rethinking LDA: Why Priors

Matter[G]BENGIO Y, SCHUURMANS D, LAFFERTY J D, et al. Advances in Neural Information Processing Systems 22 : Curran Associates, Inc., 2009 : 1973 – 1981. http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.pdf.

[83] ARORA S, GE R, HALPERN Y, et al. A Practical Algorithm for Topic Modeling with Provable Guarantees[C]. ICML'13 : Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 : JMLR.org, 2013 : II – 288. http://dl.acm.org/citation.cfm?id=3042817.3042925.

[84] BLEI D M, LAFFERTY J D. Dynamic Topic Models[C]. ICML. 2006 : 113 – 120.

[85] CHA Y, CHO J. Social-network Analysis Using Topic Models[C]. SIGIR. 2012 : 565 – 574.

[86] RUBIN T N, CHAMBERS A, SMYTH P, et al. Statistical Topic Models for Multi-label Document Classification[J]. Mach. Learn., 2012, 88(1-2) : 157 – 208.

[87] PEIRSMAN Y, HEYLEN K, GEERAERTS D. Size matters: tight and loose context definitions in English word space models[C]. Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics. 2008 : 34 – 41.

[88] HENDERSON K, ELIASSI-RAD T. Applying Latent Dirichlet Allocation to Group Discovery in Large Graphs[C]. SAC. 2009 : 1456 – 1461.

[89] STEVENS K, KEGELMEYER P, ANDRZEJEWSKI D, et al. Exploring Topic Coherence over Many Models and Many Topics[C]. EMNLP-CoNLL. 2012 : 952 – 961.

[90] WANG X, JIA Y, ZHOU B, et al. Computing Semantic Relatedness Using Chinese Wikipedia Links and Taxonomy[J]. Journal of Chinese Computer Systems, 2011, 32(11) : 2237 – 2242.

[91] FINKELSTEIN L, GABRILOVICH E, MATIAS Y, et al. Placing Search in Context: The Concept Revisited[J]. ACM Trans. Inf. Syst., 2002, 20(1) : 116 – 131.

[92] RUBENSTEIN H, GOODENOUGH J B. Contextual Correlates of Synonymy[J]. Commun. ACM, 1965, 8(10) : 627 – 633.

[93] LI W, MCCALLUM A. Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations[C]. Proceedings of the 23rd international conference on Machine learning. 2006 : 577 – 584.

[94] ISHWARAN H, RAO J S. Spike and slab variable selection: frequentist and bayesian strategies[J]. The Annals of Statistics, 2005, 33(2) : 730 – 773.

[95] WANG C, BLEI D M. Decoupling Sparsity and Smoothness in the Discrete Hierarchical

Dirichlet Process[G]. Advances in neural information processing systems. 2009 : 1982 – 1989.

[96] YIN J, WANG J. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering[C]. KDD '14 : Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA : ACM, 2014 : 233 – 242. http://doi.acm.org/10.1145/2623330.2623715.

[97] ZUBIAGA A, JI H. Harnessing Web Page Directories for Large-scale Classification of Tweets[C]. Proceedings of the 22nd international conference on World Wide Web companion. 2013 : 225 – 226.

[98] LINDSEY R V, III HEADDEN W P, STIPICEVIC M J. A Phrase-discovering Topic Model Using Hierarchical Pitman-Yor Processes[C]. EMNLP-CoNLL '12 : Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012 : 214 – 222.

[99] WANG X, MCCALLUM A, WEI X. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval[C]. ICDM : Proceedings of the 2007 Seventh IEEE International Conference on Data Mining. Washington, DC, USA : IEEE Computer Society, 2007 : 697 – 702. http://dx.doi.org/10.1109/ICDM.2007.86.

[100] GRIFFITHS T L, TENENBAUM J B, STEYVERS M. Topics in semantic representation[J]. Psychological Review, 2007, 114 : 2007.

[101] LAU J H, BALDWIN T, NEWMAN D. On Collocations and Topic Models[J]. ACM Trans. Speech Lang. Process., 2013, 10(3) : 10:1 – 10:14.

[102] EL-KISHKY A, SONG Y, WANG C, et al. Scalable Topical Phrase Mining from Text Corpora[J]. Proc. VLDB Endow., 2014, 8(3) : 305 – 316. http://dx.doi.org/10.14778/2735508.2735519.

[103] HE Y. Extracting Topical Phrases from Clinical Documents[C]. Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[104] ZUO Y, ZHAO J, XU K. Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts[J]. Knowl. Inf. Syst., 2016, 48(2) : 379 – 398. http://dx.doi.org/10.1007/s10115-015-0882-z.

[105] CHENG X, YAN X, LAN Y, et al. BTM: Topic Modeling over Short Texts[J]. IEEE Trans. Knowl. Data Eng., 2014, 26(12) : 2928 – 2941. http://dx.doi.org/10.1109/TKDE.2014.2313872.

[106] ZUO Y, WU J, ZHANG H, et al. Topic Modeling of Short Texts: A Pseudo-Document View[C]. KDD '16 : Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA : ACM, 2016 : 2105 – 2114. http://doi.acm.org/10.1145/2939672.2939880.

攻读博士学位期间取得的学术成果

发表论文：

[1] **Yuan Zuo**, Junjie Wu, Hui Zhang, Deqing Wang, Hao Lin, Fei Wang, Ke Xu, Complementary Aspect-Based Opinion Mining Across Asymmetric Collections, ICDM, 2015. (CCF B, EI)

[2] **Yuan Zuo**, Jichang Zhao, Ke Xu, Word network topic model: a simple but general solution for short and imbalanced texts, Knowledge and Information System, 48(2): 379-398, 2016. (CCF B, SCI)

[3] **Yuan Zuo**, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, Hui Xiong, Topic Modeling of Short Texts: A Pseudo-Document View, SIGKDD, 2016. (CCF A, EI)

[4] Fei Wang, Rui Liu, **Yuan Zuo**, Hui Zhang, He Zhang, Junjie Wu, Robust Word-Network Topic Model for Short Texts, ICTAI, 2016. (CCF C, EI)

[5] Hao Lin, Hengshu Zhu, **Yuan Zuo**, Chen Zhu, Junjie Wu, Hui Xiong, Collaborative Company profiling: Insights from an Employee's Perspective, AAAI, 2017. (CCF A, EI)

投稿论文：

[1] **Yuan Zuo**, Junjie Wu, Hui Zhang, Deqing Wang, Ke Xu, Complementary Aspect-based Opinion Mining, Transactions on Knowledge and Data Engineering (TKDE), under major revision. (CCF A, SCI)

致谢

看着编译成稿的毕业论文，回忆起决定转博时内心的忐忑不安以及五年博士生涯的起起伏伏，内心不禁思绪万千。不得不说，六年的硕博连读是我人生目前为止最为难忘的一段经历。在经过这个重要里程碑之际，谨向曾经给过我许多无私帮助的老师、同学以及我的家人表示最诚挚的感谢。

首先由衷感谢我的导师许可教授，本论文的选题和研究过程是在许老师的悉心指导和严格要求下完成的。许老师严谨求实的治学态度，精益求精的工作作风使我深受感染，一直并将不断地激励我以勤奋、认真、踏实的态度对待科研工作。许老师广博的学识，开阔敏捷的思维和独到的见解开拓了我的研究思路，提出了许多建设性的意见，令我受益匪浅。此外，在读博期间，许老师也对我在学习和生活方面给予了无微不至的关怀，使我能够专心学业。

特别感谢我的副导师张辉教授对我研究工作的帮助。张老师为人热情，治学严谨，致力于为学生创造良好的学习和工作环境。这些都极大地激发了我在项目、研究等工作上的热情。除此之外，张老师还会在生活中帮助我们，与我们一起参加户外活动，增进师生之间的感情，增强了同学之间的凝聚力。

特别感谢赵吉昌和吴俊杰老师对我研究工作的帮助。我的研究工作是由赵老师领进门的。从最初的实验设计，到实验结果的分析，再到最后的论文写作，赵老师一步一步教会了我如何进行学术研究。吴老师在数据挖掘领域有很高的研究造诣。吴老师言传身教的帮助我认识科研，他对于科研的态度对我影响极深，尤其是他对高质量研究的追求。

非常感谢课题组的陈林、倪江峰、李镇安、孙铭涛、马永星、陈君龙、陈勇、刘峤、赵元浩、王红升、金凌、彭升辉、王立印、张玮红、何晓楠、赵亚辉、王兴光、张伟凡、陈航、张震、蒋贤林、王飞、张文杰、张鹤、梁满庭、林剑颖、李宁、张晓鹏、李健等朋友，在平日学习和研究中给我的帮助，那些共同奋斗的日子将是我生命中最宝贵的记忆。

最后我要感谢我的父母和我的女朋友陈雪。没有你们的支持，我是很难坚持到博士论文成稿的今天的。谢谢您们。

作者简介

1989 年 6 月 24 日出生于江苏省盐城市。

2007 年 9 月考入天津理工大学计算机与通信工程学院，2011 年 7 月本科毕业并获得工学学士学位。

2011 年 9 月考入北京航空航天大学计算机学院攻读硕士学位。

2012 年 9 月硕转博攻读博士学位至今。