

上海交通大学

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

BACHELOR'S THESIS



论文题目：学术大数据中的主题演变分析

学生姓名：黄昕

学生学号：5120809026

专业：信息工程(中法合作办学)

指导教师：王新兵教授

学院(系)：上海交大-巴黎高科卓越工程师学院

学术大数据中的主题演变分析

摘要

学术资源和学术论文数量的不断增长造就了学术大数据的发展，如何在学术大数据中理解和把握学术主题的发展趋势是一个应用广泛的重要课题。学术大数据中话题演变的研究目的包括学术主题提取，主题相关论文查找，学术主题热度变化探究和学术演变趋势挖掘。在学术主题话题演变的研究中，已有的工作集中在探究话题的热度和内容随时间变化的规律，而话题间的演变趋势的探究较少。与此同时，在学术数据中，学术文献之间的引用网络是除文本信息外重要的信息，它通过将新旧文档相连的方式，在一定程度上反映了学术话题的变化趋势。如何将学术网络和主题模型结合，是个值得探究的问题。本文将学术数据中的文本信息，引用网络和论文的时间信息相结合，提出了联合主题模型，将文档的时间和引用关系融合进文档的生成过程中。通过该模型，本文对学术主题内容挖掘，重要论文推荐，学术主题热度和学术演变趋势的探究进行了建模，并构建反映领域发展趋势的学术话题地图。在实验中，模型被应用在 AAN, Arxiv HEP-PH 和 Citeseer 三个领域不同的学术数据集上，并取得了令人满意的结果。

关键词：主题模型 话题演变 学术大数据

Topic Evolution in Scholarly Big-data

ABSTRACT

The continuous growth of the number of academic papers and other resources promotes the development of scholarly big-data. How to reveal the evolution of academic themes is an important issue with wide ranges of applications. The study of topic evolution in scholarly big-data includes extracting scientific topics, searching for milestone papers, mining topic trend over time and discovering the evolution patterns among topics. While the changes in topic keywords and popularity over time have been studied before, the relations among new topics and old topics in a scientific field are rarely explored. In academic data, the citation network links papers published in different periods, which implicitly reflects how an academic field changes over time. In this sense, how to combine citation network in the study of topic evolution is an important question. In this paper, we propose a novel joint topic model, involving the citation information and time-stamps of scientific literatures into the generation process of topic model. With this model, we can discover scientific topics as well as the evolution patterns among them, which finally help to construct a topic map of a scientific field. Experiments on three datasets *AAN*, *HEP-PHY*, *Citeseer* have demonstrated that our model can effectively discover the topic evolution in scholarly big-data.

KEY WORDS: Topic Model, Topic Evolution, Scholarly Big-Data

目 录

插图索引	v
表格索引	vi
主要符号对照表	vii
第一章 绪论	1
1.1 课题背景	1
1.2 课题简介和研究现状	2
1.3 课题研究内容及目标	3
第二章 学术论文的文本建模	5
2.1 主题提取考虑与目标	5
2.2 语言模型	5
2.2.1 语言模型与统计语言模型	5
2.2.2 一元模型与混合一元模型	5
2.2.3 潜在语义分析模型	6
2.2.4 概率潜语义分析	7
2.3 LDA 语言模型	8
2.3.1 LDA 模型简介	8
2.3.2 LDA 模型推断	10
2.4 学术论文的主题提取	12
第三章 学术论文的引用网络建模	14
3.1 引用网络建模目标	14
3.2 基于社区发现的引用网络生成算法	15
3.3 基于主题模型的引用网络生成算法	16
3.3.1 Pairwise Link-LDA	16
3.3.2 混合成员模型	17
3.4 网络建模总结与分析	18
第四章 联合建模与演变表示形式	19
4.1 学术话题演变的意义和主要问题	19
4.2 话题时间的建模	20
4.3 文本, 时间和引用网络的联合建模	21

4.3.1	文本信息的建模	21
4.3.2	引用网络的建模	22
4.3.3	主题时间的建模	22
4.3.4	模型综述	22
4.3.5	模型推断	24
4.4	训练算法与编程框架设计	25
4.5	话题演变表现形式	26
第五章	联合建模在实际学术网络中的应用	27
5.1	研究学术话题演变的步骤	27
5.2	数据集收集	27
5.3	数据预处理	28
5.4	实验与结果展示	28
5.4.1	实验参数设置	28
5.4.2	文档主题提取展示	29
5.4.3	主题代表性文章	31
5.4.4	话题时间和热度展示	34
5.4.5	话题演变趋势展示	36
5.4.6	话题地图展示	38
5.5	实验结果总结	41
第六章	结语	42
6.1	课题总结	42
6.2	未来研究方向	42
附录 A	模型推断与参数估计	43
A.1	变分推断	44
A.2	参数推断	45
参考文献		47
致 谢		50
英文大摘要		51

插图索引

2-1 PLSA 模型	7
2-2 LDA 模型生成路径	9
2-3 LDA 模型	9
2-4 吉布斯采样推断下的 LDA 模型	11
2-5 主题抽取实例	12
3-1 Citeseer 学术引用网络图	14
3-2 隶属图模型	15
3-3 Pairwise Link-LDA 模型	17
3-4 Link-LDA 模型	17
4-1 话题演变实例	20
4-2 TOT 模型	21
4-3 联合建模模型	23
5-1 课题流程图	27
5-2 数据样例	29
5-3 AAN 数据集文本主题提取样例	30
5-4 HEP-PH 数据集文本主题提取样例	30
5-5 Google Scholar “Machine Translation” 搜索结果	32
5-6 Google Scholar “Congestion Avoidance Control” 搜索结果	33
5-7 Google Scholar “End to End Internet” 搜索结果	33
5-8 话题热度样例	34
5-9 AAN 数据集话题样例	35
5-10 HEP-PH 话题热度随时间变化	35
5-11 计算机网络话题热度随时间变化	36
5-12 Citeseer 数据集话题样例	37
5-13 “机器翻译” 话题演变趋势	37
5-14 HEP-PH 主题地图	38
5-15 AAN 话题地图	39
5-16 Citeseer 话题地图	40

表格索引

5-1 AAN 数据集主题 27 代表性文章	31
5-2 Citeseer 数据集主题 26 代表性文章	32

主要符号对照表

d	文档
l	引用
w	单词
t	时间
D	文档数量
L_d	文档 d 的引用长度
N_d	文档 d 的长度
K	主题个数
C	网络社区个数
T	时间段个数
V	词典中单词个数

第一章 绪论

1.1 课题背景

随着计算机的出现和发展，人类社会逐渐进入了信息化社会。在信息化社会中，资料、文件中很大一部分都被制作成便于保存和传输的电子文档，这些电子资料构成了大数据时代的基础。二十世纪以来，在网络的发展和社交网络的兴盛的背景下，人类活动在网络上的频率显著提升。大量数据也随着人类活动而产生，这一切造就了大数据时代的发展和繁荣。

大数据时代的影响也体现在学术界上。在科学和技术高速发展的今天，随着各种各样的科研活动展开，每年都会源源不断地产生来自全世界的学者、研究机构的学术数据。这些学术数据包括包括论文、书籍、专利以及技术专利等。不断庞大的学术数据的产生和积累，无论对走在科学前沿的学术界，还是对谋求创新发展的工业界，都起着极其重要的指导和参考的作用。一方面，工程师们通过了解最新到理论知识，将科学前沿的技术与工业生产相结合，促进产业技术的创新和发展。另一方面，学者通过阅读最新的学术文献，提出或拓展出新的研究课题和方向，加速各领域的进步和发展。与此同时，学术数据的传播和普及极大地增加了研究人员，开发人员之间的学术交流，从而提高了研发人员的工作效率，缩短了研究成果产生的周期。

通常概念之下，大数据具有快速发展 (Velocity)、数据格式的多样化 (Variety) 与数据量的规模大 (Volume) 的三个主要特征^[28]，同时也具有高价值，高真实度等特点。学术大数据也具有以上的主要特征。学术大数据的大规模和高速发展特征体现在它日益增长的数据储量和惊人的增长速度。例如，2010 年，微软学术¹拥有存储量超过五百万篇的学术文档记录。同时，从 1997 年至 2006 年，包括 PubMed²，Sci³等几个数据库的数据量每年有 2.7% 到 13.5% 不等的增长^[16]。除此之外，在学术数据中，很大比例的学术数据可以在网络中直接免费获取，在 2008 年到 2011 年所发表的文献中，平均 43% 的文献可以免费下载^[29]，这一事实给学者的研究工作提供了极大的便利。与此同时，学术大数据的多样性体现在资料的种类繁多，除了最常见的学术论文外、还有技术文档、专利、教学视频和讲座幻灯片。这些日益增加的学术资源，造就了近年来学术大数据的蓬勃发展。在这个背景下，网络上开始出现大大小小基于学术数据库建立的学术搜索系统，他们提供了包括文献文本、引用网络、作者信息等不同种类的学术信息。例如，Citeseer⁴的数据库包括超过三十万位作者信息、超过两百万篇文章信息以及超过五千两百万条边的学术文献引用网络。Google 从 2004 年开始，通过构建学术搜索引擎 Google Scholar⁵，向全球的学者和实验室提供海量的学术文献资源。除此之外，比较具有代表性的学术搜索系统还包括 PubMed, arXiv⁶, IEEE Digital Library⁷等。在我国，学术搜索系统的搭建和学术大数据的研究也逐渐成为一个不可忽视的方向。其中，比较具有代表性的是清华大学所开发

¹<https://academic.research.microsoft.com>

²<https://www.ncbi.nlm.nih.gov/pubmed/>

³<https://www.soci.org>

⁴<https://citeseerx.ist.psu.edu>

⁵<https://scholar.google.com>

⁶<https://arxiv.org>

⁷<https://ieeexplore.ieee.org>

的 AMiner¹ 系统。AMiner 主要利用数据挖掘数据算法和社交网络分析的技术，对学术大数据的作者，学术话题等信息进行挖掘。自 2006 年上线以来，AMiner 系统已经包含了近 8 千万学术文献数据。除此之外，国内还存在包括上海交通大学的 Acemap² 系统等。因此，学术大数据分析逐渐成为数据挖掘的领域中的重要方向。

学术大数据分析主要包括研究学术主题的演变，学术文献在引用网络中的影响力分析，学术论文重要度的排序，热门研究方向的寻找以及论文作者间的相互联系的挖掘等。其中，本课题目标是研究学术大数据中的主题演变分析，目的在于从海量学术大数据中挖掘出各领域中的主题，并得到关于主题的形成，消失和主题内容的变化趋势，即展现出一个领域话题变化和话题间关系地图。学术话题演变的研究具有重要意义。掌握学术话题的发展规律能够帮助研究人员在学术大数据中准确地把握当前的热门主题的内容，相关的文档及其发展的历史和形成过程，快速从成百万，上千万的学术文献中找到与当前学者研究最相关的文献，提高其研究的工作效率。

1.2 课题简介和研究现状

主题演变的挖掘包括以下几方面的内容。一方面是挖掘主题的内容随着时间发生变化的趋势，另一方面是了解新主题产生和旧主题的衰减的过程。在学术话题演化研究中，重要的任务就是通过挖掘文本和论文网络，得到话题的内容信息及其活跃时间，并结合并建立不同时间的话题之间的内在的联系，构建能够反映领域内话题演变和话题间相互联系的关系地图。

主题模型是挖掘文本主题的重要方法。主题模型遵循词袋模型的假设，在这个假设下，一篇文章被视为一个词的集合，词的先后顺序可以交换。在一篇文章中，每个词语的生成被看作是相互独立的过程，通过概率图模型，主题模型可以从文章的文本信息中提取相关的主题。其中，概率潜在语义分析 (Probabilistic Latent Semantic Analysis, PLSA 模型)^[15] 和 LDA 文档主题生成模型^[6] (Latent Dirichlet Allocation, LDA 模型) 是主题模型中的重要组成部分。PLSA 模型是一个非监督的生成模型，它引入了“潜在主题”的概念。在给定一篇文章后，对于该文章中的每一个词，首先以一定的概率选择对应的主题，然后以一定概率选择该主题对应的单词。LDA 模型在 PLSA 模型的基础上，在主题的分布上方增加了狄利克雷先验，从而在一定程度上克服了 PLSA 模型过拟合的缺点。在 LDA 模型被提出的近十年来，不同的学者基于 LDA 模型和 PLSA 模型，将主题模型应用于不用场景，从而形成了包括研究主题相关性的相关主题模型^[3]，研究主题层次性的分级狄利克雷模型^[12] 和针对主题随时间演化的 [1, 4, 14, 21, 25, 27] 的一系列工作。

针对话题演变所提出的主题模型主要包括 Topic Over Time 模型 (TOT 模型)^[27]，Dynamic Topic model (DTM 模型)^[4]，Continuous Dynamic Topic model (cDTM 模型)^[25]，Infinite Dynamic Topic Model (iDTM 模型)^[1]，Inheritance Topic Model (ITM 模型)^[14] 以及 Multiscale Topic Tomography (MTTM 模型)^[21]。

其中，TOT 模型^[27] 将时间因素加入模型的生成过程，在每一个单词生成的过程，作者采用 Beta 分布对主题下单词所处的时间进行建模，将文本、词和时间三者联合起来作为观测数据。通过模型训练，作者可以拟合出话题热度在时间上的变化曲线。DTM 模型^[4] 将文档集按照时间进行分段，对每个时间段的文档使用类似 LDA 模型的方法提取该时间段的主题信息；在前后时刻的主题之间，作

¹<https://cn.aminer.org>

²<https://acemap.sjtu.edu.cn>

者在不同时间段的主题—单词分布之间建立链式结构，后一时刻分布是以前一时刻的分布为均值的正态分布，并借助卡尔曼滤波的模型来捕捉不同时间段上主题的关键词变化。cDTM 模型^[25] 在 DTM 模型的基础上，考虑了时间的连续性。在建模的时候，作者并不是简单粗暴地将时间分段，而是使用布朗运动来模拟表示话题分布的参数在时间上的演化。iDTM 模型^[1] 是 DTM 模型在主题数目上也进行的延伸。在主题在时间上的出现和消亡的假设下，作者使用了分级狄利克雷过程，打破了在主题数目上的限制，实现了对文本的潜在结构进行建模。通过参数推断，iDTM 模型得到包括主题个数、主题分布，以及主题趋势的信息。ITM 模型^[14] 在模型的构建中，考虑到引用关系中所隐藏的时间顺序，提出了文章文本的生成不仅需要参考当前主题的内容，还需参考其引用文档所在年代的主题信息的假设。在文本的生成过程中，ITM 模型将文章为自发部分和继承部分。自发部分的单词遵循标准 LDA 模型的生成过程，而在继承部分中，单词来源于该文章所引用文章的主题，遵从该主题下单词分布而随机生成。MTTM 模型^[21] 在主题模型的生成过程中使用了泊松过程进行建模。模型中的泊松参数代表着词语在给定某话题上出现的期望次数。与此同时，MTTM 模型将时间层次性地分割成两个时间段，形成一个类似二叉树的层次结构。在这个树状模型中，父节点的泊松参数由其两个子节点的泊松分布按一定比例组合而成。

然而，以上针对话题演变的工作更多地强调主题的内容随着时间的变化趋势，而对主题间的演化关系研究较少。我们认为，主题的演化规律，即主题的分裂聚合的过程和新旧主题的依赖关系，是描述一个领域主题变化最直观的表现形式。与此同时，在学术文献中，由于作者在撰写文献时候总是需要引用与研究相关的文献，而且被引用到文献的发表时间在引用文章发表之前，因此文献之间的引用关系在很大程度上反映了新旧主题的演化规律。针对学术数据这一特点，[26] 将 LDA 模型中的词袋替换成文档袋，并从文档到引用网络中挖掘出文档—主题和主题—文档分布，通过构建主题间的依赖关系，挖掘出主题随时间融合、分裂、出现、消失的过程。然而，该工作忽略了文本的文本信息，仅仅从文章标题中提取关键词表示主题的内容。由于文献标题所包含的有效信息有限，这种主题提取和表现方法难以准确、完整地表现出主题的内容。与此同时，训练数据中，较新的文章被生成的次数较少，较早发表的文章被生成的次数较多，导致挖掘的主题时间相对靠前。

1.3 课题研究内容及目标

本课题的主要内容和目的是构建能够更加准确挖掘主题演化的模型，并结合已有工作和相关的算法挖掘出学术大数据中的主题演化趋势，即通过挖掘文本信息，引用关系网络信息，得到话题的相关信息，并建立话题在不同时间点之间的联系，从而构建反映领域内话题演变和话题间相互联系的关系图谱。

该模型旨在解决以下几个问题：

1. 如何提取文档的主题信息和表达主题的语义？
2. 如何寻找特定时间内的主导主题？
3. 如何表示话题热度随时间的变化趋势？
4. 如何挖掘主题之间的依赖程度以及表现主题的演化规则？

本论文主要对本课题所开展的工作和得到结果进行描述与总结，本文共分为 6 章，各章主要内容组织如下：

第一章，说明本课题的背景、目的，并讨论解决的主要问题；

第二章,介绍本课题针对学术论文主题提取的主要方法;
第三章,介绍本课题针对学术引用网络建模的意义和方法;
第四章,介绍本课题结合文本主题提取方法和网络挖掘方法所提出的模型;
第五章,对模型在学术数据集上获得的结果进行分析和讨论;
第六章,总结本课题所完成的任务,展望课题未来发展的方向。

第二章 学术论文的文本建模

2.1 主题提取考虑与目标

学术数据中包含了论文、专利、技术文档等，因此，学术大数据中绝大部分的数据以文本的形式呈现。挖掘文本中的主题信息，则是学习学术大数据中主题演变的起点和根基。主题具有一个比较宽泛的定义。在语言学家的眼中，主题往往是由一系列经常同时出现并且意义相近，用于表达特定语意一组词语组成。简而言之，主题是词的聚类。例如，描述机器学习这一主题概念的时候，人们常常会提到“概率图模型”，“深度学习”，“监督学习”，“增强学习”等一系列词语。在文本建模的角度中，主题可以由条件概率的形式表示，给定一个主题之后，我们根据一定的概率得到一个词语。数学上，这个条件概率以 $P(w|z)$ 表示。给定主题 z ， $P(w|z)$ 表示单词 w 出现的概率， $P(w|z)$ 的数值越大，则表示单词 w 与主题 z 越相关。与此同时，一篇文章可以由不同的单词描述，而同一单词也可以描述不同的主题。主题模型提出的目的，是得到文本中文本—主题分布 $P(z|d)$ 和主题—单词分布 $P(w|z)$ 。因此，在本课题中，我们采用主题模型的方法进行对学术文献的主题提取。

2.2 语言模型

2.2.1 语言模型与统计语言模型

语言模型 (Language Model) 是一用数学模型，用于描述和把握语言的内在规律。总的来说，语言模型可分为两种类别，分别是基于语法的语言模型 (Grammar-Based Language Model) 和统计语言模型 (Statistical Language Model)。基于语法的语言模型构建于语言学语法之上，由于相关的语法规则源自语言学知识，基于语法的语言模型不适用于处理海量文本的场景。统计语言模型则是基于概率统计而建立的语言模型。基于所获得的训练数据，统计语言的工作是对每个单词、句子、段落所生成的概率进行推断。

2.2.2 一元模型与混合一元模型

一元模型是最简单的统计语言模型，它基于词袋假设 (Bag of Words) 假设建立。在词袋模型中，一篇文档可以看作一个袋子，袋子里面装有词典里面的所有单词，这些单词在文档中的位置是可以交换的。对于文档而言，每个单词的生成相当于从袋子里面按照一定概率随机取出一个词。文档的生成则视为重复执行从词袋去词的过程。

对于一篇文档，它的每一个词的生成遵从同一多项分布。由于文档中的每一个词的生成过程是相互独立的，因此，文档 d 的生成过程可由以下公式表示：

$$P(d) = \prod_{i=1}^{N_d} P(w_i) \quad (2-1)$$

其中， N_d 表示文章 d 中的单词数， w_i 代表文章 d 中第 i 个单词。

一元混合模型^[22]是在一元模型的基础上建立的。与一元模型相比，一元混合模型引入了一个离散随机变量 z ，我们称之为“主题”。在一元混合模型之中，文档里单词的生成分为两个步骤：首先，在文档层面，我们一个主题 z ；然后，对于文章中的每一个词 w ，根据所选择的主题 z 和相关的条件概率 $P(w|z)$ 。在这个生成框架下，一篇文档 d 被生成的概率可由公式2-2表示：

$$P(d) = \sum_{z=1}^K P(z) \prod_{i=1}^{N_d} P(w_i|z) \quad (2-2)$$

其中， K 为主题的个数， N_d 表示文章 d 中的单词数， w_i 代表文档 d 中第 i 个单词。

在混合一元模型的假设中，每篇文档只由一个主题表示，而主题则由特征词的分布表示。然而，根据 [6] 的验证，一元混合模型在处理海量文本的情况下的效果有限。

2.2.3 潜在语义分析模型

潜在语义分析模型 (Latent Semantic Analysis, LSA 模型)，在信息检索领域也被称为 LSI 模型 (Latent Semantic Index)^[8]，于 1990 年由 Scott Deerwester, Susan T. Dumais 提出。在 LSA 模型中，单词和文档以向量的形式表示，通过向量间的距离远近关系，如计算向量与向量之间的 *cosine* 距离，我们可以得知文档间的相关程度。与传统的传统向量空间模型 (Vector Space Model) 相比，LSA 模型将词和文档映射到潜到了语义空间中，以达到了除去原始向量空间中的一些“噪音”的目的，从而提高了模型的精确度。

LSA 模型主要建立在矩阵理论中的奇异值分解 (Singular value Decomposition, SVD) 的基础上。执行 LSA 模型的时候，文本信息以单词-文本矩阵 C 表示。在这个矩阵中，每一行对应一个单词，每一列对应一个文档。矩阵中每一个元素上的数值，如第 i 行，第 j 列上的元素，表示的是单词 i 在文档 j 上的频数。然后，我们对这个矩阵进行奇艺值分解，如公式2-3所示，得到三个矩阵。其中，矩阵 T 和矩阵 D 均为正交矩阵，前者表示分解以后的单词矩阵，后者表示分解以后的文档矩阵。矩阵 S 是一个对角矩阵，矩阵对角线上的数值是奇艺值，奇艺值由上到下以逐步递减对形式呈现。

$$C = TSD^T \quad (2-3)$$

假设矩阵 S 的秩为 R ，则该矩阵有 R 个特征值。我们保留其前 K 个奇艺值，其他的清零处理。其中， $K < R$ 且 $K < \min(V, D)$ ， V 和 D 分别是单词和文档的个数。这一步骤之后，我们得到了一个新的矩阵 S' 。

如公式2-4所示，将所得到的矩阵 S' 和矩阵 T 和矩阵 D 相乘，可以重新构成新的单词-文本矩阵 C' ，所得到的矩阵被称为原单词-文本矩阵的 K -秩近似矩阵。与原来的单词-文本矩阵相比，重构的单词-文本矩阵仅保留了权重最大的 K 个维度的文本信息，从而达到了筛选文本信息中所含的噪声的目的。因此，重构的矩阵能更好的反映文本信息中的潜在语意信息。

$$C' = TS'D^T \quad (2-4)$$

然而，在处理文本信息的时候，LSA 模型存在着较大的局限性。一方面，由于 LSA 模型是以 SVD 为基础构建的，因此当文本集或者词库发生变化时，需要对单词-文本矩阵进行 SVD 计算。由于 SVD 算法的时间空间复杂度较大，因此不适合推广到研究海量数据的应用场景中。另一方面，LSA

模型在降维的过程中，只保留奇异值较大的维度。因此，在将维的过程中，文本集里词频较少但比较专业的术语容易被过滤丢失。与此同时，LSA 模型的过程也缺乏严密的统计学基础作为支撑，这也是后来的模型提高的方向。

2.2.4 概率潜语义分析

概率潜语义分析^[15](Probabilistic Latent Semantic Analysis, PLSA 模型)由 Hofmann 于 1999 年在 SIGIR 会议上提出。PLSA 模型由 LSA 模型发展而来，与 LSA 模型缺少严密的数理推理不同，PLSA 模型是一个以概率统计为基础的生成模型。

与混合一元模型中一篇文档只对应一个主题的设定不同，在 PLSA 模型中，一篇文档可以由多个主题混合而成，而每个主题都是单词上的概率分布表示，给定一个文档，每个词上方的主题生成遵循用一个多项分布。

与其他生成模型相似，PLSA 模型的设计从本质上而言是对人类撰写文章的过程的模拟。在我们撰写文章之前，基本上会执行两个步骤，首先确定文章需要包含哪几方面的主题，接着确定我们将花费多少篇幅在不同的主题上。譬如构思一篇研究学术主题演变的论文，首先，我们可能会花费 20% 的文笔对学术大数据的背景进行概述，在这一部分，我们比较倾向使用例如“大数据”，“数据容量”，“学术数据库”等词语；接着，我们可能会花费 40% 的篇幅进行描写我们针对学术主题演变所提出的模型，假如我们以主题模型为基础建立新的模型，那么在这一部分，我们可能比较倾向使用“语言模型”，“生成模型”，“概率潜语义分析”，“LDA 主题模型”，“吉布斯推断”等与主题模型息息相关的词语。在最后的 40% 篇幅中，我们可能会着重展示我们模型的效果，在这里我们可能比较倾向于“热度”，“关键词”，“主题地图”等一类直接描述学术话题演变的词语。在行文的过程中，我们之所以经常能够联想到这一些特定的词，是因为这些词语与我们所指定的主题息息相关，换句话来说，在给定的主题下，这些词被产生的概率很高。因此，一篇文档在除去介词，代词等无实际意义的词语之后，可看作多个主题的集合，而每一个主题则由与该主题相关的频率最高的一些词进行描述。

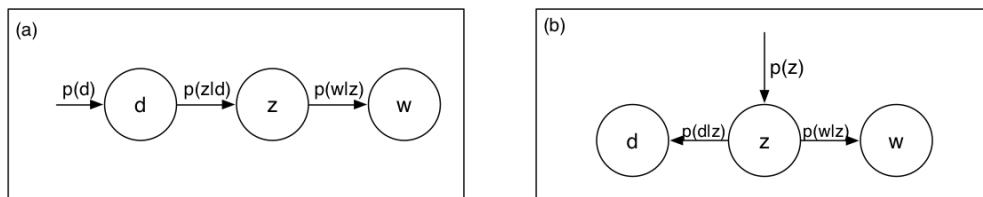


图 2-1 PLSA 模型
Fig 2-1 Probabilistic Latent Semantic Analysis

PLSA 模型的建立是对撰写文章过程的一个数学化的表示，图 2-1 表示的 PLSA 模型的概率图模型。在图 2-1 中， d 代表的是文档， z 代表主题， w 表示文档中的词语。 $P(d_i)$ 代表文档 d_i 出现的概率，这与文档的长度有关； $P(z_k|d_i)$ 代表给定文档 d_i 与主题 z_k 有关的单词被生成的概率，遵循多项分布； $P(w_j|z_k)$ 表示给定主题后，单词 w_j 被生成的概率，遵循多项分布。总而言之，对于文档中每一个词，在 PLSA 模型中的生成过程分为以下三步：

第一步：以 $P(d_i)$ 的概率生成文档 d_i ；

第二步：以 $P(z_k|d_i)$ 的概率生成主题 z_k ；

第三步：以 $P(w_j|z_k)$ 的概率生成单词 w_j 。

由于在 PLSA 模型中，每个词的生成过程相互独立，与词语在文档中的顺序无关，因此 PLSA 模型依旧符合词袋模型的假设。给定文章 d_i ，词语 w_j 被生成的概率由公式 2-5 表示。

同时，文章 d 与词语 w 的相关程度由公式 2-6 表示。

$$p(w_i|d_j) = \sum_{z=1}^K p(z|d_j)p(w_i|z) \quad (2-5)$$

$$p(w, d) = p(d) \sum_{z=1}^K p(z|d)p(w_i|z) \quad (2-6)$$

主题挖掘的目的是得到主题 z 的表现形式。在 PLSA 模型中，主题 z 属于模型中的隐藏变量。文档 d 和词语 w 属于模型中被观测到的数据。在文档 d 中，其隐含变量，即主题 z ，通过条件概率 $P(z|d)$ 生成，然后，根据主题 z 以及条件概率 $P(w|z)$ 生成单词 w 。对模型之中两组条件概率， $P(z|d)$ 和 $P(w|z)$ 的估计，是模型推断的目的。模型参数估计中采用的方法是极大似然估计，从本质上而言，参数估计的目标就是寻找到隐含变量 z ，使得生成所得的文本数据概率最大。在最大似然估计中，PLSA 模型使用期望最大化算法（Expectation Maximization Algorithm，EM 算法），求得概率空间中的局部最优解。EM 算法将在本章 2.3.2 节进行介绍。

与 LSA 模型相同，PLSA 模型考虑了文档中同义词和多义词的问题，在此基础上，采用了 EM 算法来，而且相对于 LSA 模型有着坚实的统计学基础，这也使得 PLSA 模型属于更加完善的模型。

然而，PLSA 模型也具有一定的局限性：

- PLSA 模型的学习过程中，采用的是 EM 算法进行学习，与此同时，模型产生的参数数量为 $KD + VK$ ，随着文档数量增多，参数数量线性增加，EM 算法的学习结果往往会产生过拟合的效果，使模型缺乏延展性和预测性。
- PLSA 模型的生成过程只包含了训练集中的文档。对训练集以外的新的文档而言，模型无法得到他们的先验，这使得 PLSA 模型无法从训练集以内扩展到训练集以外。

综上所述，PLSA 模型具有过拟合且难以扩展的局限性，但 PLSA 模型生成过程中对隐含变量的描述的思想，是值得借鉴的。

2.3 LDA 语言模型

2.3.1 LDA 模型简介

LDA 模型^[6]（Latent Dirichlet Allocation）是一种文档主题生成模型，由 David Blei, Andrew Ng 和 Michael Jordan 在 2003 年的 JRML¹ 上提出。目前，LDA 模型及其相关的模型是信息处理领域中最重要的模型之一。

LDA 模型的建立基于撰写文档时的基本假设，文档由一系列主题构成，主题下方是词语的分布。一篇文档可以包含多个主题，而同一个单词可以和不同的主题相关联。在这个假设下，LDA 由一个

¹ Latent Dirichlet Allocation 的会议论文版本为 [5]

三层的贝叶斯概率图模型构建，这三层由上到下分别文档，主题和单词。与前文所提及的统计语言模型相一致，LDA 遵循了词袋假设，一篇文档在除去无实际意义的停词之后，被抽象称为一个词频向量，与此同时，文档之下存在一个由主题所构成的概率分布，而每一个主题下方是一个单词所构成的一个概率分布。

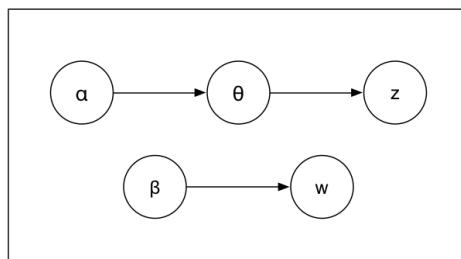


图 2-2 LDA 模型生成路径

Fig 2-2 generation schema of Latent Dirichlet Allocation

与 PLSA 模型等语言模型一致，LDA 模型是一个生成模型。对于一篇文档 d ，在生成文档的每一个词的时候，我们首先以一个多项式分布生成一个主题 z ，该多项分布记为 θ ，例如， $P(z|\theta)$ 表示给定 θ 时，主题 z 生成的概率分布。接着，我们根据所选择的主题 z 所对应的主题—单词分布 β ，从词典中的 V 个单词里生成单词 w ，概率为 $P(w|z, \beta)$ 。在词典上方，LDA 模型添加了超参数为 α 的狄利克雷 (Dirichlet) 分布作为文档—主题分布的先验。如图2-2所示，LDA 的生成过程分为两层，分别是主题的生成和单词的生成。LDA 总的生成过程可以分为以下步骤：

- 对于每一篇文档，首先由超参数为 α 的狄利克雷先验分布求主题 z 的多项式分布参数 θ ，即 $\theta \sim Dirichlet(\alpha)$ 。其中 $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ ，分别表示文档下第 k 个主题 z_k 被生成的概率。
- 对于文档里的每一个词，首先根据 θ 生成一个主题 z_n 。即 $z_n \sim Multi(\theta)$ 。然后根据被选定的主题 z_n ，从分布 $P(w_n|z_n, \beta)$ 中选择一个词 w_n 。文档的生成可被视为重复词语生成的过程。

LDA 模型可由一个贝叶斯概率图模型表示，其图模型如图2-3表示。

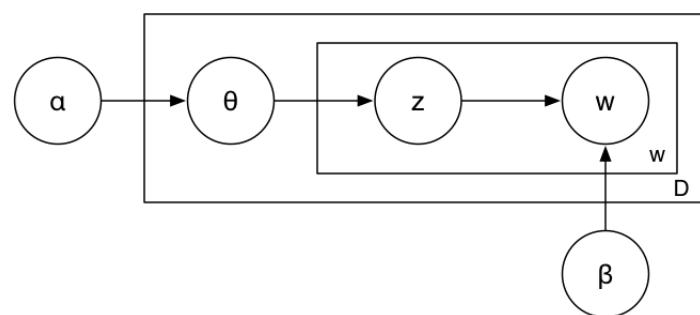


图 2-3 LDA 模型

Fig 2-3 graph of Latent Dirichlet Allocation

图中， α 为狄利克雷先验分布的超参数， θ 是一个维度为 k 的向量，代表着文档-主题分布。 β 是一个 $K \times V$ 的矩阵。矩阵中 β_{ij} 代表了第 i 个主题下生成单词 j 的概率， w 表示的是文章中所生成的

单词。在 LDA 模型中，一篇文章被生成的概率可由公式 2-7 表示。

$$P(d|\alpha, \beta) = \int P(\theta|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_n=1}^K P(z_n|\theta) P(w_n|z_n, \beta) \right) d\theta \quad (2-7)$$

与 PLSA 模型相比，LDA 模型在前者的基础上在文档—主题层引入了狄利克雷 (Dirichlet) 分布先验，对模型进行了再一层的贝叶斯封装。因此，LDA 模型与 PLSA 模型最大的区别在于主题的生成上。在 PLSA 模型中，所有文档的主题分布都看作模型的参数，随着训练文档数量的增加，模型参数的数量也随之线性增加，训练结果过拟合的问题也随之而来。而 LDA 模型引入了超参数 α 对主题对生成进行了建模，因此无论文档数量多少，LDA 模型的最外层只有一个超参数影响主题的生成。超参数的引入使的 LDA 模型较 PLSA 模型相比有以下两点提高：

- LDA 模型降低了 PLSA 模型过拟合的程度，鲁棒性得到了增强；
- LDA 模型可以应用于训练集未包含的文档中。

这两点的提高，使得 LDA 模型比 PLSA 模型更加适用于大规模文本信息的处理当中。

2.3.2 LDA 模型推断

LDA 模型的模型推断主要目的是得到文档—主题分布和主题—单词分布。常用的推断方法包括吉布斯采样推断和变分 EM 算法。

2.3.2.1 吉布斯采样推断

吉布斯采样算法是马尔可夫链蒙特卡洛方法 (Markov Chain Monte Carlo, MCMC) 的实现方式，算法的思想在于构建一条最终平稳收敛与目标概率函数的马尔可夫链。由于在 LDA 模型中，先验分布狄利克雷分布，与目标分布多项分布是共轭分布，因此，吉布斯采样算法可以应用在 LDA 模型的推断上。在论文 [13] 中，作者介绍了吉布斯采样运用于 LDA 模型推断过程的方法。应用吉布斯采样时，如图2-4所示，主题—单词分布上层添加了类似于文档—主题分布的狄利克雷先验，狄利克雷分布的超参数为 β ，采样目标概率函数在 LDA 模型中为文档—主题分布和主题—单词分布。在每个循环中，在文档中根据马尔可夫链现有状态所对应的分布情况抽取所有变量值，从而转换马尔可夫链中分布进入下一状态，这一步骤重复执行，直到状态收敛至马尔可夫链的平稳状态。LDA 模型的吉布斯采样算法由算法由算法2-1所示。

算法 2-1 Gibbs sampling for LDA

1. 初始化：对每篇文档的每一个词 w , 随机赋予主题 z
 2. 扫描整个文档集，根据 [13] 里的 Gibbs Sampling 公式，对每篇文档的每一个词 w , 重新赋予主题 z
 3. 重复步骤 2 直到收敛
 4. 输出 θ 和 ϕ
-

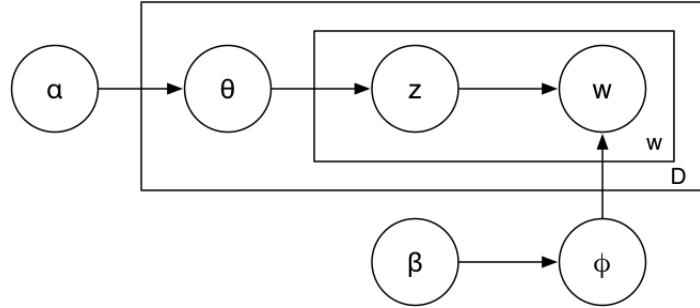


图 2-4 吉布斯采样推断下的 LDA 模型
Fig 2-4 Latent Dirichlet Allocation for Gibbs Sampling

2.3.2.2 EM 算法与变分 EM 算法

最大期望 (EM) 算法^[9] 在机器学习中常用的一种推断算法。EM 算法的目的是寻找参数最大似然估计或者最大后验估计。对于模型中存在无法观测的隐藏变量的概率模型，如 PLSA 模型和 LDA 模型中的主题，EM 算法是推断模型适合使用的算法。

EM 算法分为交替进行的 E 步和 M 步：

- 在 E 步中，根据隐藏变量的当前估计值，计算其似然值；
- 在 M 步中，寻找最大化在 E 步上求得的似然值所对应的参数的值。

与 EM 算法一致，变分 EM 推断的思路来源于数理统计中参数估计常用的极大后验概率方法，即求解生成文本数据概率最大时所对应的参数。然而，在 LDA 模型的三层贝叶斯网络中，参数 θ 和 β 之间存在程度很高的耦合，因此参数的估计无法通过一般的极大后验概率的方法求解。针对这种情况，模型推断采用的方法是贝叶斯统计中变分推断的方法。这种办法的核心思想在于构造一个能够分离变量的参数去无限逼近原本需要求解的参数，最后得出最优解下对应的参数解值。

在 LDA 模型中，所需求解的参数分别是主题分布 $P(z)$ 和单词分布 $P(w)$ 。衡量构造分布和原分布间距离的方式，通常采用的是 Kullback-Leibler divergence，KL-divergence 的定义由公式 2-8 所示。

$$KL(P||Q) = \sum P(i) \log\left(\frac{P(i)}{Q(i)}\right) \quad (2-8)$$

在推断的过程中，构造一个分布 q ，如公式 2-9a 所示。该分布与隐变量 θ 和 z 的联合分布的 KL-divergence 如公式 2-9b 所示。

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod q(z|\phi) \quad (2-9a)$$

$$KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) = \int \sum_z q(\theta, z|\gamma, \phi) \log\left(\frac{q(\theta, z|\gamma, \phi)}{p(\theta, z|w, \alpha, \beta)}\right) d\theta \quad (2-9b)$$

利用贝叶斯公式，我们可以将公式 2-9b 改写成公式 2-10a：

$$\log(p(w|\alpha, \beta)) = L(\gamma, \phi|\theta, z) + KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) \quad (2-10a)$$

$$L(\gamma, \phi | \theta, z) = \int \sum_z q(\theta, z | \gamma, \phi) \log\left(\frac{q(\theta, z | \gamma, \phi)}{p(\theta, z | w, \alpha, \beta)}\right) d\theta \quad (2-10b)$$

其中， L 记作分布 $p(w|\alpha, \beta)$ 的下界。给定超参数的情况下， $p(w|\alpha, \beta)$ 可以看作是一个定值。在这个情况下，极大化下界的过程，也是极小化分布 $q(\theta, z | \gamma, \phi)$ 与目标分布 $p(\theta, z | w, \alpha, \beta)$ 之间 kl-divergence 的过程。这一步属于变分 EM 算法的 E 步，而 EM 算法的 M 步则是保持隐藏变量不变，求解极大化对数似然函数 $p(w|\alpha, \beta)$ 所对应的 α 和 β 的参数值。在最优化求解的过程中，使用的方法是拉格朗日乘数法，EM 算法的流程由算法2-2所示。

算法 2-2 Variational inference for LDA

1. 初始化全局参数 α 和 β 以及局部参数 γ 和 ϕ
- 2.E 步：根据 [6] 中公式更新参数 γ 和 ϕ
- 3.M 步：根据 [6] 中公式更新参数 α 和 β
4. 重复步骤 2 和步骤 3 直到参数收敛

2.4 学术论文的主题提取

LDA 模型对文档主题的提取和表达方式主要是通过模型中文档—主题分布和主题—单词分布来体现。我们以论文 *One Sense Per Discourse*^[11] 为例。该在文章中，作者报告了两个新的词义消歧系统，将系统在双语训练集 *Canadian Hansards* 和单语训练集 *Roget's Thesaurus and Grolier's Encyclopedia* 上进行训练，并发现了存在同一完整语句的多义词通常情况下表达相同的语义。通过 LDA 模型，我们得到了主题篇文章的 θ 参数，并通过 β 参数抽取了该篇文章的主题信息。

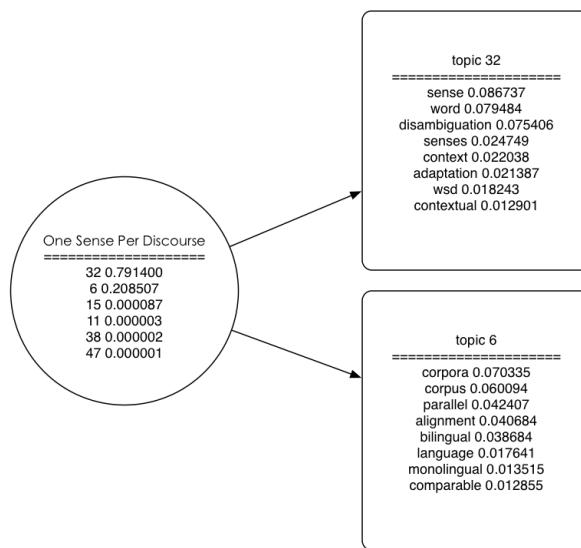


图 2-5 ”One Sense Per Discourse” 主题抽取
Fig 2-5 topic extraction of “One Sense Per Discourse”

图2-5是我们对该文章的主题抽取的结果。该文章有两个主要主题，分别是主题 32 和主题 6。在

占据主导的主题 32 中，比较具有代表性的词为 “sense”，“word”，“disambiguation” 等，是词义消歧对应的英语单词。而在少数主题 6 包含 “bilingual”，“monolingual” 两词，对应着文章中实验所采用的单语训练样本和双语训练样本。

从上述例子可以得知，LDA 模型在主题意义抽取上有着良好的性能，因此，LDA 模型的生成过程也是我们研究学术数据话题演变在主题内容获得方面的一个重要参考。

第三章 学术论文的引用网络建模

3.1 引用网络建模目标

本文在第二章主要讨论了通过主题模型从学术论文的文本信息中提取主题的方法。提取主题信息是学习学术大数据中主题演变的第一步。与此同时，我们注意到学者在撰写学术文献的时候，往往会展开时间较前但主题相关的学术论文、专利或者技术报告作为学术论文的参考文献。在主题模型的假设中，学术论文在去除无意义的停词过后，可以被看作主题的集合。因此，引用的过程可以被看作一个建立主题与主题之间联系的过程。除此之外，在学术论文引用的过程中，被引用的论文总是发表在新的论文之前，而不存在学术论文引用未来发表的论文的可能性。因此，论文之间的引用关系隐含着话题随时间变化。简言之，学术论文之间的应用关系在一定程度上反映旧主题和新主题之间的演化关系。

因此，在提取学术文档主题的基础上，研究学术大数据中主题演变的下一步就是挖掘学术论文之间引用网络的信息。这一部分的目的是从学术文档之间到引用关系的角度分析主题之间的依赖程度与变化趋势，包括主题内容的变化，以及多个主题聚合形成单个主题，或者分裂为不同的发展方向的过程。学术论文的引用网络从本质上讲是一个有向图，因此，在引用网络的建模上主要应用的是图的生成算法。

社区发现问题是网络分析中的重要问题，其的目的是在网络中发现网络社区的存在。社区指的是一个节点的集合，处于同一社区的节点具有某种相同或者相近的性质。在学术网络中，社区由若干篇文章构成。

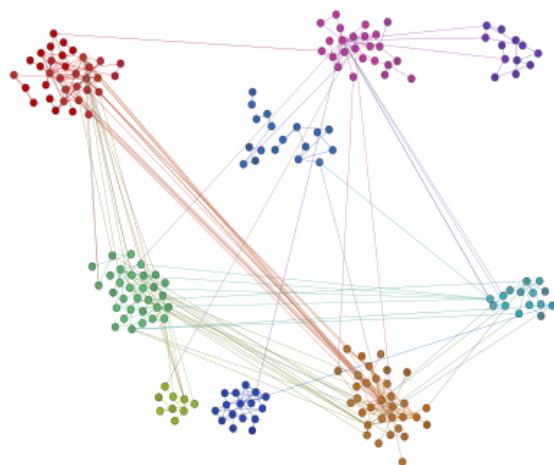


图 3-1 Citeseer 学术引用网络图
Fig 3-1 Citation network of Citeseer

图3-1是我们从 *Citeseer* 数据集中随机抽出部分文章所得到的论文引用网络图。在图中，每一个节点代表了一篇学术文献，文献间的引用关系则由图中的边表示。我们使用 openord^[17] 算法进行布局，在这个布局下，相互之间边较为密集的节点将会聚合在一起，形成网络社区。在这个 *Citeseer* 引用网络中，尽管因为随机采样导致网络不完善，我们依旧可以发现文章在引用网络中聚合成若干个社区。我们分别用九个不同颜色标注局部网络图里处于不同社区之中的节点。通过观察这九个社区间点与点之间的边。我们可以发现另一个现象：处于相同社区的文章成批引用另一社区的文章。如图中被橙色标定的节点中绝大部分节点与红色社区的节点相连，同样的情况还出现在黄色社区和绿色社区，绿色社区和红色社区之间。

引用网络图的这两个现象反映了学术引用网络之间所存在的两个特征。首先，在社区网络之中，社区是存在的，他由若干个学术论文组成。在同一社区中，文章之间往往存在相对密集的引用边。其次，在学术网络中，社区与社区之间存在相互联系的，处于同于社区不同文章很大程度会同时引用另一个社区中的文章。由于相互引用的文章在主题层面有不同程度上的相同或者相近，社区间文章相互引用这一现象在一定程度上可以体现出主题之间通过引用网络产生关联的现象。因此，学术网络中社区发现是挖掘话题演变规律的一个重要方面。

3.2 基于社区发现的引用网络生成算法

隶属图模型^[30](Affiliation Graphical Model, AGM 模型) 由 Jaewon Yang 等人在 2012 年的 ICDM 会议提出。AGM 模型是用于生成网络的模型。在调查 6 个不同网络包括四个不同的社交网络以及亚马逊用户购买网络和 DBLP 学术网络之后，作者发现网络社区存在较高的重叠性，这个重叠性表现在一个节点存在很高的概率隶属于不同的社团。AGM 模型的建立基于以上这一现象上，其生成过程由图3-2所示：图中的 a , b 分别代表网络中的节点， u 代表网络中的社区。参数 F_{au} 和 F_{bu} 分别代表着

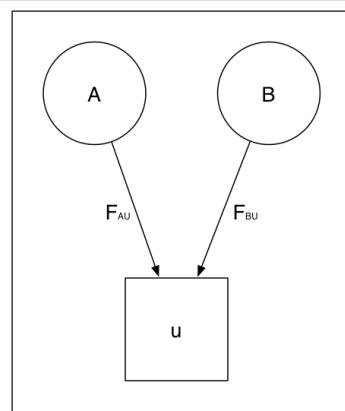


图 3-2 隶属图模型
Fig 3-2 Affiliation graphical model

节点 a 和节点 b 属于节点 u 的强度。 a , b 两点之间属于同一社区 u 的强度之积越大，则 a , b 两点通过社区 u 所产生联系的概率越大。

在 [31] 之中，作者将节点的特征考虑进图的生成过程和社区发现的过程中。通过逻辑回归建模，

作者可以从训练数据中推断出网络中的节点属于社区的强度以及社区与节点特征之间的关系。

AGM 模型能够适应于大规模的网络，并可以发现网络中高度存在的社区信息。然而，AGM 模型的初始化过程使得模型的应用具有局限性。

AGM 模型的初始化是一个极其复杂的过程。对于社区数目的选择，AGM 模型参考了 [2] 的方法，首先，我们将网络中的节点分成 20% 和 80% 的两部分，通过调整社区数目，我们不断训练 80% 部分的点构成的网络，并计算关于 20% 节点点似然值。所选择的社区数量对应的是所得似然值最大的情况。对于强度参数 F 的初始化设置上，我们首先计算网络中每一个节点的阻抗，当社区数量为 k 的时候，我们选择阻抗最高的 k 个点，这 k 的点与每一个社区之间的强度参数 F 设为 1，其余值赋为零。对于其他的点，强度矩阵赋为零矩阵。这种初始化使得 AGM 模型属于从局部开始延伸的社区发现模型。训练结果对初始状况依赖性很大。同时，这一性质也会导致在最终结果中，只有相对较小的一部分可以寻找到所隶属的社区。在这个情况下，AGM 模型并不适用于需要得到大部分节点社区信息的场景。

3.3 基于主题模型的引用网络生成算法

除了基于社区发现的引用网络生成算法之外，主题模型在网络生成过程，尤其是对学术引用网络中的生成过程，也有着广泛的应用。其中比较代表性工作的包括 [10, 20, 19]。

3.3.1 Pairwise Link-LDA

Pairwise Link-LDA 模型^[20] 是 Ramesh Nallapati 在 2008 年 SIGKDD 会议上提出。Pairwise Link-LDA 模型是一个基于 LDA 模型的生成模型，除了考虑了文档的词信息，该模型在学术论文的引用网络上也进行了建模。

在引用网络的生成过程中，Pairwise Link-LDA 结合了混合隶属度随机块模型^[2](Mixed Membership Stochastic Block models, MMSB 模型)。MMSB 模型由 Airoldi 提出，在这个模型中，对于网络的每个节点，首先指定该节点对每个社团的隶属度，即节点属于社团的概率；其次，对于每对节点，分别生成两个节点隶属的社团以及节点由于所属社团而存在联系的概率。

Pairwise Link-LDA 模型有机的结合了 LDA 模型和 MMSB 模型，其概率图模型表示形式见图3-3，在 Pairwise Link-LDA 模型之中，社团的概念由文档的主题替代，其中矩阵表示的是主题引用矩阵，矩阵里第 i 行第 j 列的数值表示包含主题 i 的文章引用主题 j 的概率。在文档文本的生成上，Pairwise Link-LDA 遵循 LDA 模型中文档—主题，主题—单词的生成路径。

总的来说，Pairwise Link-LDA 模型结合了主题模型和网络生成模型的特点，在建模的过程中考虑了学术文献的文本信息和引用信息，同时，主题引用矩阵在本质上也反映了主题间通过引用关系演变的现象。然而，Pairwise Link-LDA 模型未能克服 MMSB 模型的局限性，在生成文章间引用网络时，需要考虑文本集里所有的文献。这一特点，使得 Pairwise Link-LDA 模型在处理大规模文献数据时候，存在极大的空间复杂度。这使得 Pairwise Link-LDA 模型不适合处理海量学术大数据资料。与此同时，在同一文档中的词语和引用连接的生成过程中，都需要经过文档—主题的生成过程。在文档数目较大的情况下，潜在链接的数量将大于单词的数量，这也导致了从文本中提取主题时主题的质量较低。

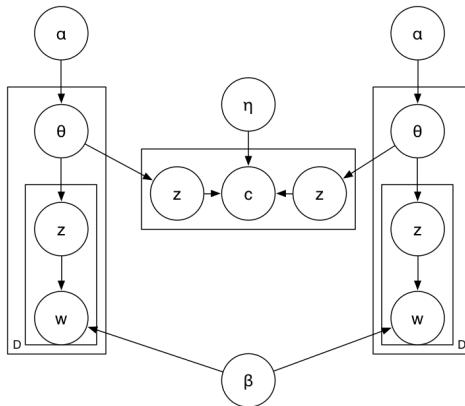


图 3-3 Pairwise Link-LDA 模型

Fig 3-3 Pairwise Link-LDA

3.3.2 混合成员模型

混合成员模型^[10](Mixed Membership Models)由Elena Erosheva于2004年提出。混合成员模型又叫Link-LDA模型，也是一个基于学术文档的单词信息以及文本的引用网络信息所建立的模型。

Link-LDA模型建立在LDA模型的基础上。在Link-LDA模型模型的生成过程中，对每篇文档词信息，Link-LDA模型依照与LDA模型相同的生成路径生成，即先按照文档—主题分布生成主题，再按照主题—单词分布生成单词。与LDA模型相比，Link-LDA模型在生成的过程中加入了文献集合中的文献之间相互引用的引用网络信息。类比于文档中单词的生成过程，在引用网络生成的过程中，Link-LDA模型使用了相似的方法：对于每一篇需要被引用的论文，首先按照文档—主题分布选择主题，接着按照主题—文档分布生成被引用的文档。从生成过程的本质上来看，一篇文档视作由词汇和参考两个部分，参考中的每条引用类比于文档里的单词，而整个文档库则类比于单词生成过程的词库。

Link-LDA的概率图模型如图3-4所示。

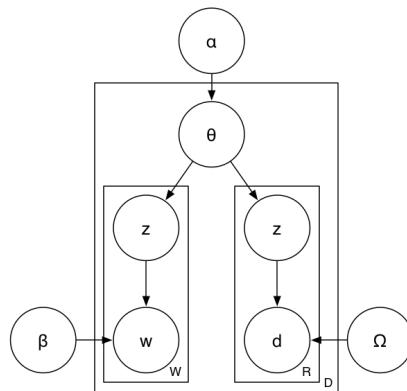


图 3-4 Link-LDA 模型

Fig 3-4 Link-LDA

Link-LDA 模型通过模仿人类撰写学术论文的两个步骤，撰写词语和引用参考文献对文本信息和引用信息进行了建模。然而与 Pairwise Link-LDA 模型相同，这里的主题既是单词上方的分布，也是文档上方的分布，也导致了从文本中提取主题时用单词所表示的主题的质量较低。

3.4 网络建模总结与分析

在3.2节和3.3节中，我们介绍和分析了基于社区发现的网络生成模型 AGM 模型和基于主题模型的学术网络生成模型 Pairwise Link-LDA 模型和 Link-LDA 模型。除此之外，基于主题模型的学术网络生成模型还包括基于 Link-LDA 模型建立的 Link-PLSA-LDA 模型^[19]。AGM 模型适用于大规模的网络图的生成，但是其初始化的方式决定该模型属于局部起步的社区发现模型，这也导致了模型中参数覆盖面不全的特点。基于主题模型的学术网络生成模型中，虽然同时考虑了学术文献的文本信息和引用信息，但是社区的含义由主题取代，这导致了主题之下包括了文章和单词两个变量，较普通 LDA 模型相比，这些模型所提取出来的主题质量相对较低。但是，基于主题模型的学术网络生成模型在生成图的过程中提供了模拟引用过程的启发性思想，同时，在 Pairwise Link-LDA 模型中的 η 参数体现了主题间通过引用的转化关系，这一系列特征都释放了对研究学术演变的积极信号。因此，在本课题中针对学术网络的建模方面，主要在基于主题模型的学术网络生成模型进行改进。

第四章 联合建模与演变表示形式

在第二章和第三章中，我们分别讨论了如何通过主题模型从文本数据中提取主题信息以及如何对文档引用网络建模。通过以上两部分的建模方法，我们能过挖掘出学术大数据中文本信息中暗含的主题，以及论文在引用网络中的位置和所在的社区。基于以上两部分的主要思想，本章讲讨论如何学习和挖掘学术大数据中的话题演变趋势。

4.1 学术话题演变的意义和主要问题

与已有工作，包括 DTM 模型^[4]，ITM 模型^[14]，cDTM 模型^[25] 等不同，本课题的研究目标不局限于研究话题内容随时间改变，还包括对探究话题之间的演变关系，和话题的演变结构的探究。在由 X. Wang 在 2013 年于 SIGKDD 会议发表的论文 [26] 中，作者讨论了研究学术话题演变的意义和需要解决的几个问题。

在学术文献和学术数据数量快速增长的今天，快速把握学术话题的内容和发展趋势是提高学者研究效率的重要手段。对于刚刚进入新的研究领域的学者而言，如果缺乏对该领域主要话题的了解和话题演变趋势的把握，该学者很容易在海量论文之前感到不知所措。相反，在了解话题产生，发展，分化的过程，学者可以比较高效地通过该领域中反映不同话题的重要文章，快速地了解新的主题，同时寻找值得研究的课题。

总的而言，研究学术话题演变的研究集中于解决以下问题。首先，在主题层面，我们需要知道和了解主题到内容，即主题包括哪些重要的关键词？同时，我们需要得到最能代表该主题的文章，即如何寻找代表性的文献？在主题之上，我们希望得知主题的热门程度，那一些主题是其他学者比较关注的主题？以及这一些学术话题是如何产生和发展，它们是否来源于旧主题，是否在当下有其他多主题替代？

回答以上问题的答案分别是寻找话题的内容，话题下文章的分布，话题的热度分布，以及话题间的发展关系。在论文 [26] 中，作者通过建模逐一给出了解答。

在模型层面，作者主要使用的是 LDA 主题模型。与运用于文本信息上的传统 LDA 模型不同的是，作者将 LDA 应用在了引用网络的构建上，并命名该模型为 Citation-LDA 模型（以下简称 C-LDA 模型）。与传统的 LDA 模型相同的是，C-LDA 模型是一个概率生成模型，它模拟的是撰写学术论文时引用参考文献的过程，以达到生成引用网络的目的。在引用一篇参考其他文献的时候，作者首先决定一个与需要参考的主题，接着根据所选主题，选择一篇需要引用的参考文献。一篇文档的引用是通过重复完成选取主题及其被引用论文这一行为完成的。通过吉布斯采样推断算法，基于论文对引用信息作为训练数据，模型可以学习得出文档—话题，话题—文档两个多项分布。

在 C-LDA 模型中所得到的话题实质上是一个文档的聚类，本质上与网络中的社区意义相同，并不涉及文档的词信息。在提取主题内容中，作者采用了直接的方法，即统计单词在主题下出现的期望来表示话题的语义特征。对于其他信息，包括话题下文章的分布，话题的热度分布，以及话题间的发展关系，作者也通过模型所得到的文档—主题分布和主题—文档分布的组合给出了解答。其中，

话题下重要文章的定义直接按照话题—文档分布得出，话题的热度由话题在某一时间被生成到概率表示，话题之间的转换概率由话题之下文档相互引用概率所表示。

该文章定义了学术话题演化所需要解答的问题，并给出了相对合理的解答，同时，较为简单的模型也使得该思想能够引用与大规模数据的分析中。然而，模型在研究主题内容的时候，采取的是比较简单的手法，并没有进行深入的挖掘，导致了模型的预测性不足。同时，提取词频的方法所得到的词语往往是常用词，因此，主题提取的过程容易忽略一些专业性的词语，造成内容的缺失。

4.2 话题时间的建模

直观上，话题具有时间的属性。例如，在谈论机器学习的时候，我们常常联想到深度学习或者神经网络。而十年前，我们可能更倾向谈论支持向量机等其他内容。这说明话题是会随着时间变化而变化。我们对 AAN 数据集^[23]里的 20989 篇学术论文通过 LDA 模型提取 70 个话题，通过统计每个话题在每年的生成到概率，我们得到了话题热度随时间变化的趋势图。

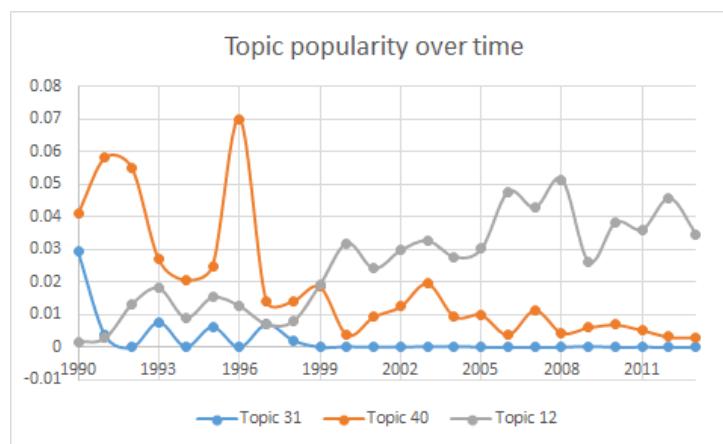


图 4-1 话题变化趋势
Fig 4-1 Topic popularity over time

图 4-1 我们随机抽取话题 12, 话题 31, 话题 40 并描绘出这三个话题的热度随着时间的变化趋势。我们发现每个话题在每个时间段上的热度是不同的。以话题 12 为例，该话题的热词是“model”, “language”, “probabilistic” 即概率语言模型，该主题与 PLSA 模型和 LDA 模型息息相关，因此，我们发现该话题在 PLSA 模型发表的 1999 年后，以及 LDA 模型发表的 2003 年后都处于一个发展的阶段。

以上发现说明了话题具有时间因素，而在对话题的建模的时候，考虑时间对话题的影响是十分必要的。

[14] 与 [27] 是主题模型与话题时间相结合的例子。[14] 主要考虑写文章时候旧时间段的主题对词语生成影响，而 [27] 主要的目的是把时间因素加入了文章的生成过程中，一方面，考虑了主题的实时性，提高主题的可观测行，另一方面则是得到了话题热度随着时间的变化趋势。对话题演变的趋势进行了建模。[27] 的模型又称为 Topic Over Time 模型 (TOT 模型)。在 LDA 模型的基础上，TOT 模型充分考虑了话题在不同时间点上的热度不同，在文档的每个单词的生成过程中考虑了该时间段与此相关主题的热度对词语生成概率的影响，在生成单词的同时还会生成主题下单词出现的时间点。

生成时间点的分布遵循 β 分布。TOT 模型的概率图表示如图4-2。

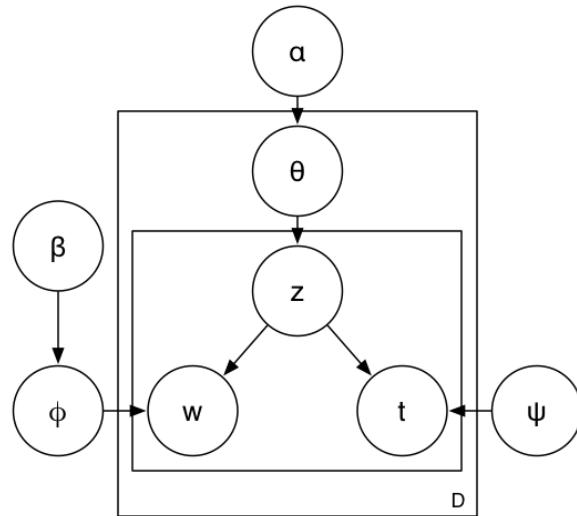


图 4-2 TOT 模型
Fig 4-2 Topic over time

我们注意到，相比于 LDA 模型，TOT 模型在生成过程中增加了单词的时间点生成，并增加了 ψ 矩阵。 ψ 矩阵是一个 $K \times V$ 的矩阵，其中每一行代表着一个主题，每一列对应着一个单词。矩阵中第 i 行第 j 列对应着的是在主题为 i 的情况下生成单词 j 时的时间 β 分步的参数。 ψ 矩阵的引入对话题，词语的生成在时间上进行了限制，也建立了话题和时间之间的联系。与此同时，我们也注意到 TOT 模型对时间建模方法与 LDA 模型相比增加了 $K \times V$ 相互独立的参数，这也导致了 TOT 具有以下局限性：

- 一方面，过多的参数在有限的数据量下会造成训练不完整的结果，使拟合的效果大打折扣，
- 另一方面，参数的增多也增加了训练的成本，这在数据量较大的情况下影响十分显著。

4.3 文本、时间和引用网络的联合建模

基于 LDA 语义模型和 C-LDA 模型在网络生成上的应用，结合 Link-LDA 模型联合建模的思想以及 TOT 模型对主题在时间上的建模，本课题建立了新的联合模型，将学术文档的词信息，文档的时间信息和引用网络信息相结合，以回答研究学术数据中话题演变所需回答的问题。

4.3.1 文本信息的建模

在2.4节中，我们分析了 LDA 模型在文本信息的主题提取上的性能。由于 LDA 模型在文本主题的提取工作上有着良好的效果，新的联合模型在文本信息的提取上继承 LDA 模型的生成方法。

4.3.2 引用网络的建模

在以前针对文本和引用网络的联合建模的工作中，包括 Link-LDA 模型^[10], Pairwise Link-Lda 模型^[20] 以及 Link-PLSA-LDA 模型^[19] 中话题, 即词的聚类, 以及网络中的社区即文档到聚类被视为同一个概念。在这个背景下, 话题既是单词的聚类, 又是文本的聚类。这种将两种不同性质聚类混为一谈的方法会导致话题通过其下方单词反映时质量偏低。我们分别使用 LDA 模型和 C-LDA 模型在 AAN 数据集上进行训练。在分析两个模型训练后所得到的参数, 我们发现, 一篇文章所属的社区与该文章中主题有很大相关性。以论文 *Lightly Supervised Learning of Procedural Dialog Systems*^[24] 为例, 该文章是关于 *Dialog Systems* 文章, 其提取的主题为主题 16, 主要的热词包括 “user”, “dialog”, “system”, “interactive”, 其所在的社区为社区 40, 该社区代表性文章有 “*Can We Translate Letters?*”, “*INeATS: Interactive Multi-Document Summarization*”, “*Crowdsourcing the evaluation of a domain-adapted named entity recognition system*”, “*Learning to Adapt to Unknown Users: Referring Expression Generation in Spoken Dialogue Systems*”, “*Hybrid Approach to User Intention Modeling for Dialog Simulation*”, “*A Flexible Framework For Developing Mixed-Initiative Dialog Systems*” 所包含的文章的主题可以由主题 16 中点关键词形容。

文档的话题与引用网络的社区相关联的原因可以由学者撰写文献过程的步骤解释。当我们撰写论文的时候, 我们首先填充的是文档的正文部分, 在写每一个词, 我们先选择文档表达的主题, 再在主题下选择单词这是 LDA 主题模型所模拟的过程。在整理论文的参考文献的时候, 我们同样先选择一个论文社团, 再从社团下选择文档作为参考文献。然而, 在引用的过程, 所引用的论文常常是与主题相关的。例如, 我们需要写一篇关于主题模型的论文, 这篇论文包括的主题如主题模型, 推断算法, 概率图模型等, 在我们引用文档的时候, 我们会引用与以上主题相关的论文, 如提出 PLSA 模型的 [15] 的提出 LDA 模型的 [6] 等。这一系列论文本身相互之间就存在引用关系, 在引用网络中, 这些论文往往存在于同一社区之中。换句话来说, 话题可以派生出学术引用网络中的网络社区。

因此, 在本课题所提出的联合建模中, 我们假设文章的主题和引用网络的社区属于不同的概念, 并将这两种不同的类之间存在的联系关系考虑进建模的过程中, 在模型中, 我们引入一个转移参数, 联系两种不同的类。

4.3.3 主题时间的建模

在生成主题时间的建模过程中, 我们参照 TOT 模型的思想, 在生成每个单词的同时, 生成一个与时间相同的观测量, 以此把主题随时间化的趋势考虑进单词的生成过程中。与 TOT 模型相比, 我们针对 TOT 参数过多的缺点, 在时间点生成过程做出了变动。首先, 我们把时间与主题直接相连, 而不是如 TOT 模型中把单词作为时间和主题联系的媒介。其次, 我们根据学术文章发表时间周期性的特点, 将时间分成固定程度的时间段, 给定一个话题, 所对应的时间遵循多项分布。而矩阵 ψ 则是主题—时间矩阵, 其元素 ψ_{ij} 表示了第 i 个主题下生成的单词所在文章发表于第 j 个时间段的概率。在这种生成方式下, 词语生成部分新增的参数数量为 $V \times T$ 较 TOT 模型大大减少。

4.3.4 模型综述

结合4.3.1节, 4.3.2节和4.3.3节的讨论, 我们提出新的主题模型。

与前文所介绍的概率语言模型相同，本课题所使用的联合建模模型也属于生成模型。在模型中，一篇文档的生成过程分为以下几步，首先，对每篇文章，生成参数 θ ，这是文档—主题分布；然后，在文章的下方，是文章中的单词和引用的生成过程：

- 对于文章的每一个词，首先根据文档—主题分布选取一个主题 z ，再根据所选择的主题 z 及其相关的主题—单词分布选择相应的单词，在生成单词的同时，还会按照主题—时间分布生成一个时间观测量。
- 对于文章中的每一个引用，首先根据文档—主题分布选择一个主题 z ，再根据所选择的主题 z 及其相关的主题—社区分布选择一个网络社区 c ，最后根据所选择的网络社区及其相关的社区—文档分布选取被引用的论文。

在以上所形容的生成过程中，本模型将论文的单词和引用链接的产生的过程是相对独立的过程，这使得主题和网络社区的聚合过程中受到较小的影响因素。与此同时，模型将文字，主题，社区，文档四个变量相连，以任意一个变量为起点，都能够找到一条路径，生成其他所有变量。这一特性可以帮助我们学习学术话题演变，并回答 [26] 所提及的问题。

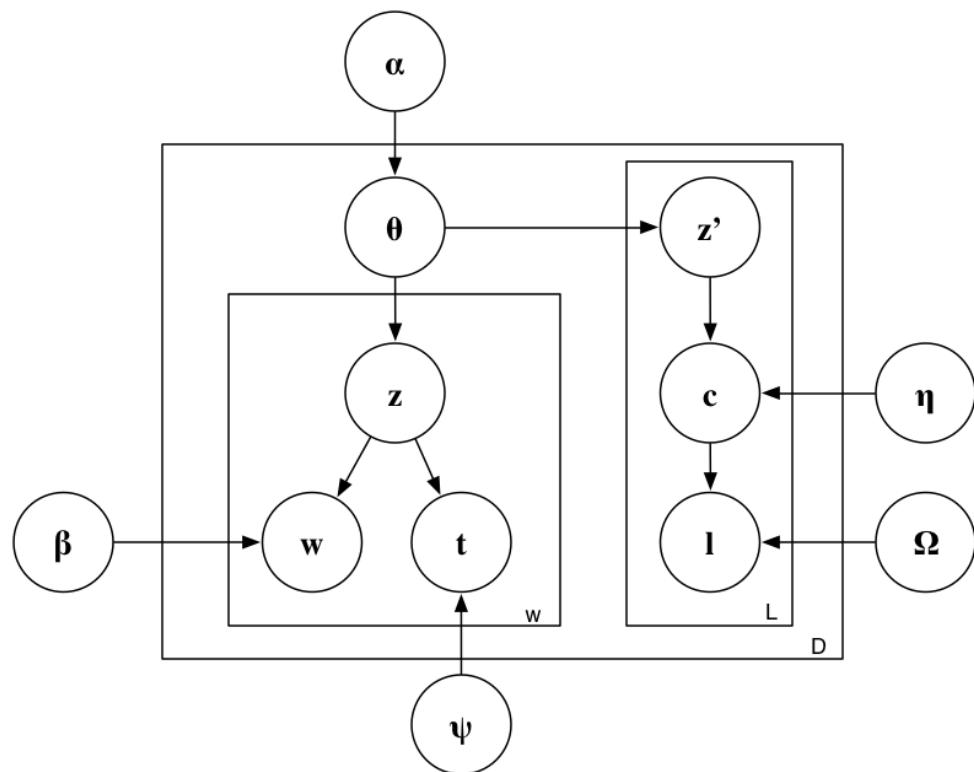


图 4-3 联合建模模型

Fig 4-3 Joint topic model

新的联合建模主题模型的概率图模型如图4-3所示。其中， α 为狄利克里分布的参数， β 为主题—单词分布， ψ 为主题—时间分布， θ 为文档—主题分布， z, z' 为主题变量， w, t 分别为单词变量和该单词的时间观测量点， η 为主题—社区分布， Ω 为社区—文档分布， c, l 分别代表网络社区，引用边。公式4-1为该模型的联合分布。

$$\begin{aligned}
 P(w, c, t | \alpha, \beta, \psi, \Omega, \eta) &= \int P(\theta | \alpha) \prod_{n=1}^{N_d} \sum_z P(z | \theta) P(w_n | z, \beta) P(t_n | z, \psi) \\
 &\times \prod_{l=1}^{N_t} \sum_{z'} P(z' | \theta) \sum_c P(c | z', \eta) P(l_l | c, \Omega) d\theta
 \end{aligned} \tag{4-1}$$

4.3.5 模型推断

从数学上，模型的推导可以简化为一个最大似然参数估计问题。即求解相关的隐含变量，使得模型的后验概率最大。

在本模型中，采用的推断公式为变分推断算法，与 2.3.2.2 节介绍的方法一致，其核心思想是构造分布 q 似以逼近并求解出所需的文档—主题分布，主题—单词分布，主题—社区分布，社区—文档分布。在这里，分布 q 如公式4-2所示。

$$q(\theta_d, z_d, z'_d, c_d | \gamma_d, \phi_d, \lambda_d, \sigma_d) = q(\theta | \gamma) \prod_{n=1}^{N_d} q(z_n | \phi_n) \prod_{l=1}^{L_d} q(z'_l | \lambda_l) \prod_{l=1}^{L_d} q(z'_l | \sigma_l) \tag{4-2}$$

在变分 EM 的框架下，公式4-1可改写为公式4-3

$$\ln(P(w, c, t | \alpha, \beta, \psi, \Omega, \eta)) = L(q) + KL(q || p) \tag{4-3}$$

其中，

$$L(q) = \int \sum_z \sum_{z'} \sum_c q(\theta_d, z_d, z'_d, c_d) \ln\left(\frac{P(w, c, t, \theta_d, z_d, z'_d, c_d | \alpha, \beta, \psi, \Omega, \eta)}{q(\theta_d, z_d, z'_d, c_d)}\right) \tag{4-4a}$$

$$KL(q || p) = \int \sum_z \sum_{z'} \sum_c q(\theta_d, z_d, z'_d, c_d) \ln\left(\frac{q(\theta_d, z_d, z'_d, c_d)}{P(\theta_d, z_d, z'_d, c_d | w, c, t, \alpha, \beta, \psi, \Omega, \eta)}\right) \tag{4-4b}$$

通过拉格朗日乘数法求解极值，我们可以得出模型中隐含变量以及参数的求解，结果由公式4-5a至公式4-5h所示。

$$\phi_{dnk} \propto \beta_{kw_{dn}} \psi_{kt_d} \exp(\Psi(\gamma_{dk})) \tag{4-5a}$$

$$\gamma_{dk} = \alpha + \sum_{n=1}^{N_d} \phi_{dnk} + \sum_{l=1}^{L_d} \lambda_{dlk} \tag{4-5b}$$

$$\lambda_{dlk} \propto \exp(\Psi(\gamma_{dk}) + \sum_{c=1}^C \sigma_{dlc} \ln(\eta_{kc})) \tag{4-5c}$$

$$\sigma_{dlc} \propto \Omega_{cl_{dl}} \exp\left(\sum_{k=1}^K \lambda_{dlk} \ln(\eta_{kc})\right) \tag{4-5d}$$

$$\beta_{kv} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} w_{dn}^v \phi_{dnk} \quad (4-5e)$$

$$\psi_{kt} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} w_d^t \phi_{dnk} \quad (4-5f)$$

$$\eta_{kc} \propto \sum_{d=1}^D \sum_{l=1}^{L_d} \sigma_{dlc} \lambda_{dlk} \quad (4-5g)$$

$$\Omega_{cp} \propto \sum_{d=1}^D \sum_{l=1}^{L_d} c_{dl}^p \sigma_{dlc} \quad (4-5h)$$

其中 $\Psi()$ 是 *Digamma* 函数。 w_{dn} 是指示变量，当 $w_{dn}=v$ 时， $w_{dn}^v=1$ 否则 $w_{dn}^v=0$ ；同理，当 $c_{dt}=p$ 时， $c_{dt}^p=1$ 否则为零，当文档 d 的时间为 t 的时候， $w_d^t=1$ ，否则为零。模型具体的推断过程将在附录中呈现。此外，参数 α 无法求出解析解，在这里我们借鉴 LDA 模型 [6] 中使用的 Newton-Raphson 算法在迭代过程中逐步更新。

4.4 训练算法与编程框架设计

模型推断的框架采用传统的变分 EM 算法相似，主题部分分为似然估计的 E 步和极大步求解的 M 步，模型的训练算法由算法4-3所示。

算法 4-3 Variational inference for Joint topic model

1. 初始化全局参数 $\alpha, \beta, \psi, \eta$ 和 Ω 以及局部参数 γ, ϕ, λ 和 σ
 2. E 步：根据公式4-5a到公式4-5d更新参数 γ, ϕ, λ 和 σ
 3. M 步：根据公式4-5e到公式4-5h以及 [6] 中 Newton-Raphson 算法更新参数 $\alpha, \beta, \psi, \eta$ 和 Ω
 4. 重复步骤 2 和步骤 3 直到参数收敛
-

编程的代码分为数据的读取，训练和输出三个部分。在程序中，我们定义了两个类，论文集和文档。论文集类代表的文档的总集合，其成员变量包括指向每篇论文的指针，论文数量，词典中词的个数。文档类的成员变量包括论文的单词数，每一个单词信息，引用数和指向每一篇被引用论文的指针。

模型的训练部分是代码的核心部分由四个被封装的函数构成，分别是函数 `run_em()`，函数 `doc_e_step()`，函数 `lda_inference()`，以及函数 `lda_mle()`；函数 `run_em()` 执行的是 EM 算法的整体工作，调用了函数 `doc_e_step()`，函数 `lda_inference()`，以及函数 `lda_mle()`。其中，函数 `doc_e_step()` 用于将每一篇文章的单词信息和引用信息传递至函数 `lda_inference()` 中，函数 `lda_inference()` 对应算法的 E 步，即更新隐含变量 γ, ϕ, λ 和 σ 并计算新的变量下的似然值，函数 `lda_mle()` 的作用对应着算法的 M 步，其任务是更新参数 $\alpha, \beta, \psi, \eta$ 和 Ω 。

4.5 话题演变表现形式

经过对所取得的数据进行训练，我们可以获得隐含在文本信息和文档引用网络中的五个不同分布，即文档—主题分布，主题—单词分布，主题—时间分布，主题—社区分布，社区—文档分布，通过这五个分布，本节讲对如何回答话题演变的五个问题进行讨论和分析。

首先，在提取文档话题上，我们可以通过主题—单词分布进行表示，话题在模型中被视为一个单词之上的变量，话题的内容以单词对概率分布表示。这种表示方法与 LDA 模型在话题关键词提取时的方法是一致的。

在提取主题内容之后，遵循相同的思想，我们在寻找反映话题的文档时候依照话题—文档的分布。这种方法与 C-LDA 模型的构建方法一致，本质上，主题—文档下方概率较大的文章属于给定主题，在引用过程中常常选择的文档。然而由于在本模型中不存在直接的主题—文档分布，因此，我们通过主题—社区分布与社区—文档分布对该问题的解答进行构建。构建的方法参照公式4-6。

$$P(d|z) = \sum_{c=1}^C \eta_{zc} \Omega_{cd} \quad (4-6)$$

对主题对热门程度的表示，我们可以用模型中文档—话题分布进行表示。给定一篇文章，在选择主题的过程中，所选择的主题概率越大则说明该主题更热门，给定时间后，该时间点上主题的热度则由该时间内选择主题的概率期望表示。数学化的形式由公式4-7表示。

$$P(z = k|t) = \sum_{d=1}^D \mathbf{1}_{d_t=t} \frac{\# < z_w = k >}{\sum_{i=1}^K \# < z_w = i >} \quad (4-7)$$

话题演变规律的挖掘分为两个部分，话题时间的确定和话题演变关系的挖掘。前者的目的是寻找主题较热门的年代，在这里，话题热度随时间的变化趋势由模型中的 ψ 参数表示。话题演变的趋势中重要的任务是寻找话题与话题之间通过引用网络产生的联系，以及衡量联系的强度。在本模型中，由于引用网络的生成由文档—话题—社区—文档路径生成，同理，话题与话题之间联系关系可以由话题—社区—文档—话题这一路径生成。数学化的公式由公式4-8表示。

$$P(z = k|z' = k') = \sum_{d=1}^D P(d|z' = k') P(z = k|d) \quad (4-8)$$

第五章 联合建模在实际学术网络中的应用

5.1 研究学术话题演变的步骤

在第四章中，我们主要介绍了研究学术话题演变所使用的模型及其推断算法和解答研究话题演变重要问题的数学化的方法。在本章之中，我们将模型应用在实际的学术大数据网络中，挖掘其中学术话题演变的发展规律。总的来说，挖掘学术数据的流程可以分为以下 4 步，分别是数据的收集，数据的预处理，模型训练，训练结果可视化。本章工作的流程图如图 5-1 所示。

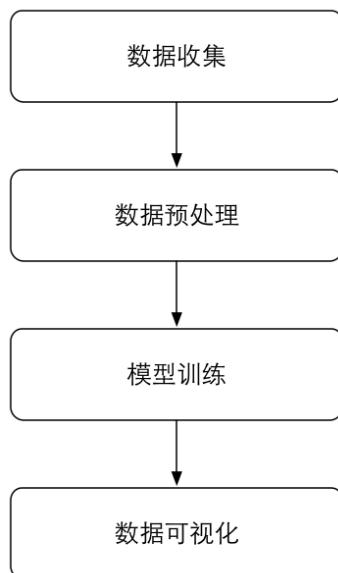


图 5-1 课题流程图

Fig 5-1 Flowchart of the project

5.2 数据集收集

本课题所涉及的模型将主要应用在三个开源的学术数据集，分别是 *AAN* 数据集，*HEP-PH* 数据集和 *Citeseer* 数据集。

AAN 数据集^[23] 全称为 *ACL Anthology Network dataset*，是一个开源的学术数据集。数据集内的文章主要涵盖了自然语言处理领域中的期刊文章和会议文章。*AAN* 数据集涵盖了文档的标题，正文，引用，发表时间等信息。*AAN* 数据集是一个封闭的数据集，数据集里所有论文的引用都发生在数据集内部。在由 *AAN* 数据集里文档构成的引用网络中，存在着 20989 个节点，即文章的个数，和 125934 条边。*AAN* 数据集是一个相对稠密的数据集，数据集中每篇文档平均大约有六条引用信息，除此之外，*AAN* 数据集的文档主要来源单一的学术领域，因此，从 *AAN* 数据集里抽取的主题专业性较强，且主题之间的相似度比较大。

*Arxiv HEP-PH*¹数据集全称为 *high energy physics phenomenology citation graph*。在 2003 年的 KDD Cup 中由斯坦福大学的 SNAP 实验室²发布。数据集内的文章主要包含高能物理领域的期刊文章和会议文章。*HEP-PH* 数据集涵盖了文档的标题，正文，引用，发表时间等信息在，由 *HEP-PH* 数据集里文档构成的引用网络中，存在着 34,546 个节点，即文章的个数，和 421,578 条边。*HEP-PH* 数据集是一个相对稠密的数据集，数据集中每篇文档平均大约有十二条引用信息，除此之外，*HEP-PH* 数据集的文档主要来源单一的学术领域，因此，从 *HEP-PH* 数据集里抽取的主题专业性更强，且主题之间的相似度更大。

*Citeseer*³是一个由宾夕法尼亚州立大学开发的电子图书馆系统，该系统内的学术文献主要包括的是计算机和信息科学领域的学术文献。在本次应用中，我们收集了 2006 年前所发布的 *Citeseer* 开源数据集，数据集包含每篇文档的文字信息，标题信息，引用信息，发表时间和作者。*Citeseer* 数据集是一个封闭的数据集，数据集里所有论文的引用都发生在数据集内部在由 *Citeseer* 数据集所构成的引用网络中，包含着 716800 个节点，代表着 716800 篇文章，和 1760574 条引用边。在本节的实验中，我们去掉了引用网络中较为孤立的节点，抽取出包含 51208 篇文章的子数据集。

5.3 数据预处理

实验前预处理的目的有两个，一方面是将网络中获取的原始数据更改成符合模型代码可读的格式，另一方面是去除原始数据的噪音，包括没有实际语义的停词、乱码、数字和符号等。同时所有单词需要统一大小写。以上工作由 *python* 代码完成。其中在提取停词的过程中，我们调用了 *python* 的 *nltk* 库，*nltk(natural language toolkit)* 是 *python* 的自然语言处理工具包，其中的英语停词库包括了 127 个停用词。

处理后的数据样式如图5–2所示。在保存文本信息的文档中，第一行是该数据集中所包含的文档的个数，随后的每一行是文章的词信息。在保存引用信息的文档中，第一行代表一篇文章，每行的第一个数字是文章的 *id*，随后的数字是该文章引用文献的 *id*。

5.4 实验与结果展示

5.4.1 实验参数设置

模型的训练参数设置如下：

- 对于 *AAN* 数据集，我们设置主题数 $K=70$ ；社区数 $C=70$ ；超参数 α 的初始值为 0.1；函数 *lda_inference()* 的收敛标准为 0.000001；全局的收敛标准为 0.0001。
- 对于 *HEP-PH* 数据集，我们设置主题数 $K=50$ ；社区数 $C=20$ ；超参数 α 的初始值为 0.1；函数 *lda_inference()* 的收敛标准为 0.000001；全局的收敛标准为 0.0001。
- 对于 *Citeseer* 数据集，我们设置主题数 $K=90$ ；社区数 $C=90$ ；超参数 α 的初始值为 0.1；函数 *lda_inference()* 的收敛标准为 0.000001；全局的收敛标准为 0.0001。

¹<http://arxiv.org/archive/hep-ph>

²<http://snap.stanford.edu/>

³<http://citeseerx.ist.psu.edu/>

29555

study algebraic aspects kontsevich integrals generating functions intersection theory moduli space review derivation virasoro kdv constraints intersection numbers kontsevich integral main theorem expansion characters schur functions proof first part theorem grassmannians kdv matrix airy equation virasoro highest weight conditions genus expansion singular behaviour equation generalization higher degree potentials quantum deformed algebra shown kinematical symmetry harmonic chain whose spacing given deformation parameter phonons symmetries well multiphonon processes derived quantum group structure inhomogeneous quantum groups thus proposed kinematical invariance discrete systems

0 3636 11652 14091 14546 15423 16447

1 5106 5113 5125 5292 8323 8574 8649 9124 11682 14062 14172 14545 14970

14999 15044 15347 16447 16837

2 5125 5290 7172 8143 8574 10173 12699 13549 14360 14587 15136

3 2314 2315 8366 10135 13276 13288 14004 14005 14545 14587 15423 16466

16690 16946 17012

图 5–2 数据样例

Fig 5–2 Example of training data

5.4.2 文档主题提取展示

与 LDA 模型主题提取的方法一致，本课题所提出的联合模型在主题提取的过程中主要参照的是模型的文档-主题分布和主题-单词分布。

图5–3展示的是 AAN 数据集中论文 *Statistical Machine Translation with Local Language Models*^[18] 的主题提取结果。我们从文章的摘要提取专业术语作为图上方的方框内容。下方左圆框为我们提取的该文章的文档-主题分布，下方右方框为主题 27 的主题-单词分布。通过文档的标题以及摘要关键词，我们可以了解到本文章是关于 *Machine Translation* 和 *Statistical Language Models* 的文章。从主题-单词分布中，我们发现主题 27 在文章中占据绝对主导的地位。从主题-单词分布中，我们发现该主题下出现比较频繁的关键词包括 “*translation*” , “*system*” , “*language*” , “*machine*” , “*model*” , “*dialogue*” , “*systems*” , “*statistical*” 。

图5–7展示的是 HEP-PH 数据集中论文 *Quantum hair on black holes*^[7] 的主题提取结果。我们从文章的摘要提取专业术语作为图左边的方框内容。中间圆框为我们提取的该文章的文档-主题分布，右方框为主题 6 和主题 35 的主题-单词分布。通过文档的标题以及摘要关键词，我们可以了解到本文章是关于涉及量子理论和黑洞的文章。从主题-单词分布中，我们发现主题 6 在文章中占据主导的地位，其次是主题 35。从主题-单词分布中，我们发现该主题 6 下出现比较频繁的关键词包括 “*theory*” , “*field*” , “*quantum*” , “*string*” , 这个主题主要阐述的是量子力学相关主题，主题 35 下出现比较频繁的关键词包括 “*black hole*” , “*model*” , “*dilaton*” , “*solutions*” , 这个主题主要阐述的是黑洞的相关主题。

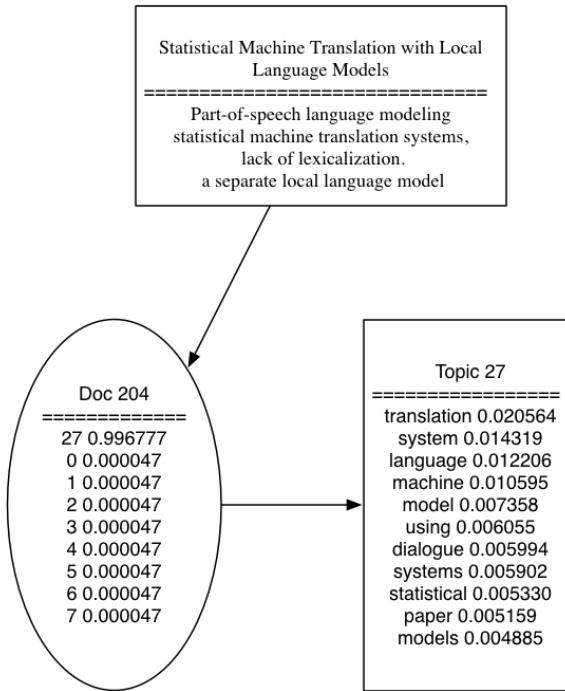


图 5-3 AAN 数据集文本主题提取样例
Fig 5-3 Example of topic extraction of AAN Dataset

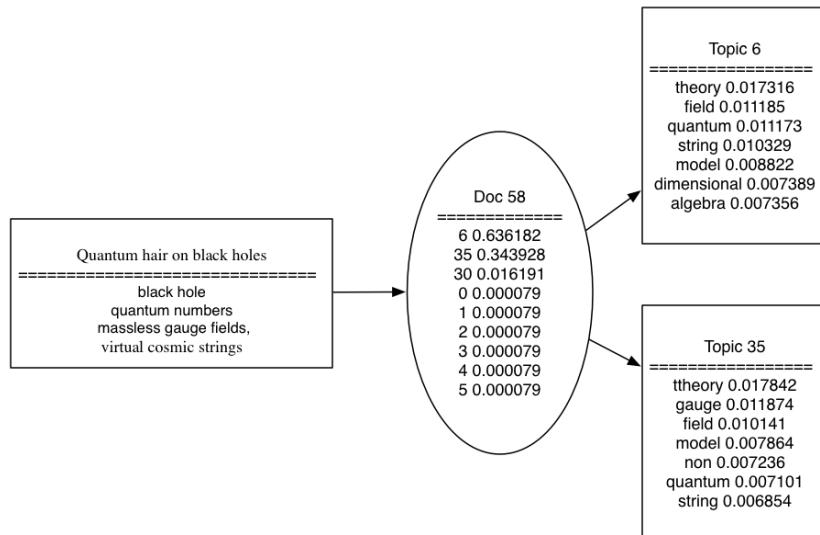


图 5-4 HEP-PH 数据集文本主题提取样例
Fig 5-4 Example of topic extraction of HEP-PH Dataset

从分布的情况和摘要的关键词对比来看，模型在文档主题提取能力上表现良好。

5.4.3 主题代表性文章

在调研一个新的领域的时候，我们常常需要寻找关于领域最具有代表性的文章。这些文章往往在这个领域具有深远的影响力，使得作者在撰写同领域的论文时候常常会引用该文章。例如，在主题模型领域中，最具有代表性的文章之一为 David Blei 发表的 *Latent Dirichlet Allocation*。在我们模型中，文档的因相互引用的关系构成了网络社区，因此由主题—社区—文档路径构成的主题—文档分布可以体现出对于给定一主题，通过该主题引用文章的概率。在寻找主题代表性文章中，主题—文档分布为我们所参考的指标，其表达式由公式4–6所示

表 5-1 AAN 数据集主题 27 代表性文章

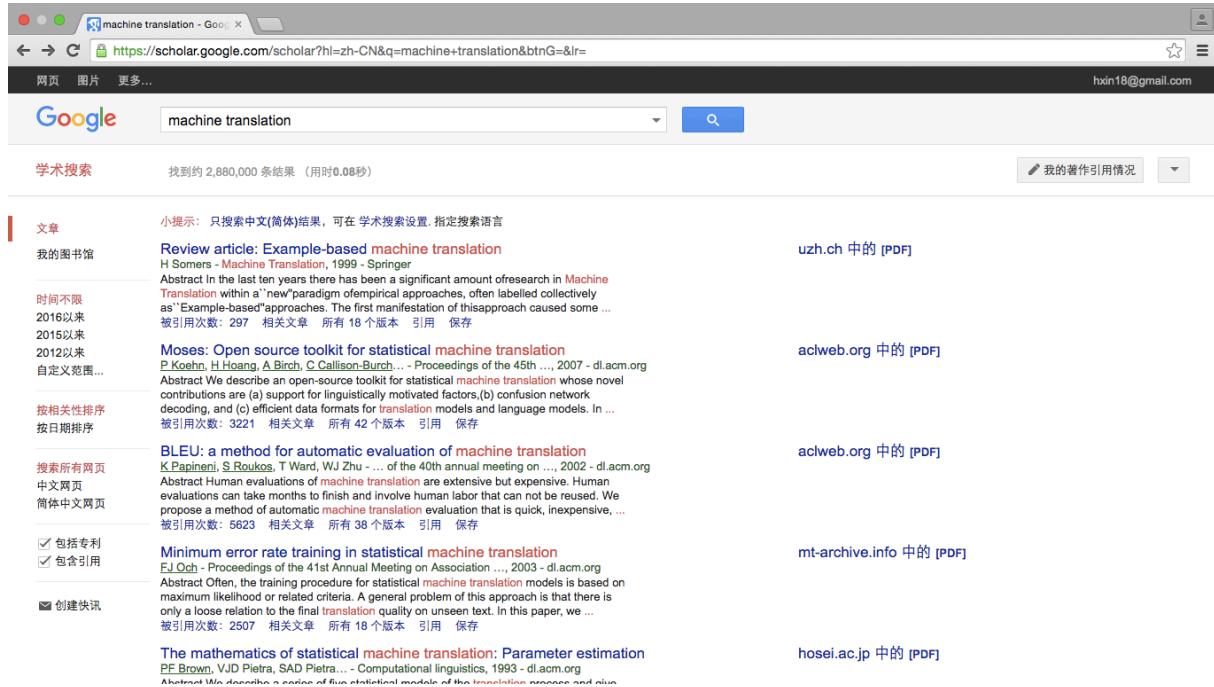
Table 5-1 Milestone papers for topic 27

Milestone Papers	Topic-Docs(%)
Bleu: A Method For Automatic Evaluation Of Machine Translation	0.031088
Moses: Open Source Toolkit for Statistical Machine Translation	0.022306
Minimum Error Rate Training In Statistical Machine Translation	0.02034
Statistical Phrase-Based Translation	0.015655
A Systematic Comparison Of Various Statistical Alignment Models	0.013819
Hierarchical Phrase-Based Translation	0.011883
The Mathematics Of Statistical Machine Translation: Parameter Estimation	0.009457
Statistical Significance Tests For Machine Translation Evaluation	0.008196
A Hierarchical Phrase-Based Model For Statistical Machine Translation	0.007706
METEOR: An Automatic Metric For MT Evaluation With Improved Correlation With Human Judgements	0.005975

表 5-1 展示的是关于 AAN 数据集中主题 27 相关的文章中的前 10 篇文章，由图5-3可知，主题 27 是关于 “Machine Translation” 的，由图表看出，所列出的文章均为 “Machine Translation” 相关的文章。在与 Google Scholar 中对 “Machine Translation” 的排名前五搜索对比中，我们发现，除去 Google Scholar 返回结果第一篇文章未存在于 AAN 数据集外，其余四篇文章，“Bleu: A Method For Automatic Evaluation Of Machine Translation”，“Moses: Open Source Toolkit for Statistical Machine Translation”，“Minimum Error Rate Training In Statistical Machine Translation” 和 “The Mathematics Of Statistical Machine Translation: Parameter Estimation” 均包含在我们所列出文章列表中。

表 5-2 展示的是关于 Citeseer 数据集中主题 26 相关的文章中的前 10 篇文章，主题 26 的词云由图5-12(a)所示，该主题是关于 “Congestion Control” 和 “End to End Internet” 的话题，这两个关键词共同形容端对端拥挤控制这一计算机网络主题。我们分别在 Google Scholar 搜索与端对端拥挤控制相关的 “Congestion Avoidance Control” 和 “End to End Internet” 这两组词语，通过对比结果，我们发现，在 “Congestion Avoidance Control” 和 “End to End Internet” 这两次搜索中，Google Scholar 返回的结果中，排名靠前的文章，除去不包含在我们训练集中的文章外，均在我们给出的推荐列表中。

从通过主题—文档分布所得文档和 Google Scholar 搜索结果对比来看，模型在寻在对主题具有代表性的文档上有一定的表现能力。



The screenshot shows a Google Scholar search results page for the query "machine translation". The search bar at the top contains "machine translation". Below the search bar, it says "找到约 2,880,000 条结果 (用时 0.08 秒)". On the left, there is a sidebar with filters: "文章", "我的图书馆", "时间不限", "2016以来", "2015以来", "2012以来", "自定义范围...", "按相关性排序", "按日期排序", "搜索所有网页", "中文网页", "简体中文网页", "包括专利", "包含引用", and "创建快讯". The main results list includes several papers with their titles, abstracts, and download links (e.g., [PDF] or [aclweb.org]). One result is from "uzh.ch" and another from "aclweb.org". Other results are from "hosei.ac.jp" and "mt-archive.info". The results are presented in a standard list format with some basic filtering applied.

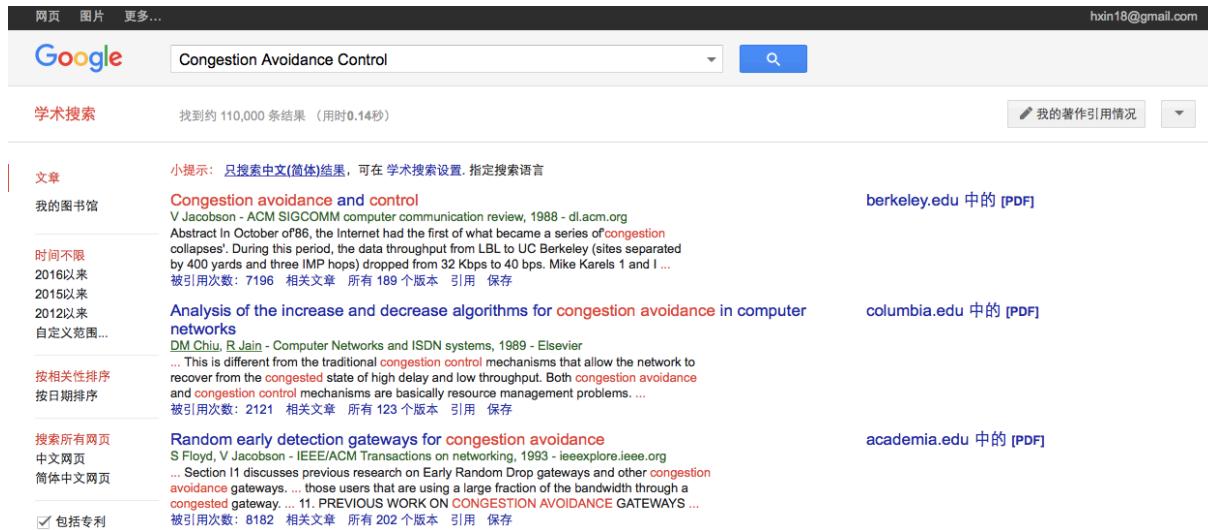
图 5-5 Google Scholar “Machine Translation” 搜索结果

Fig 5-5 Search result of “Machine Translation” on Google Scholar

表 5-2 Citeseer 数据集主题 26 代表性文章

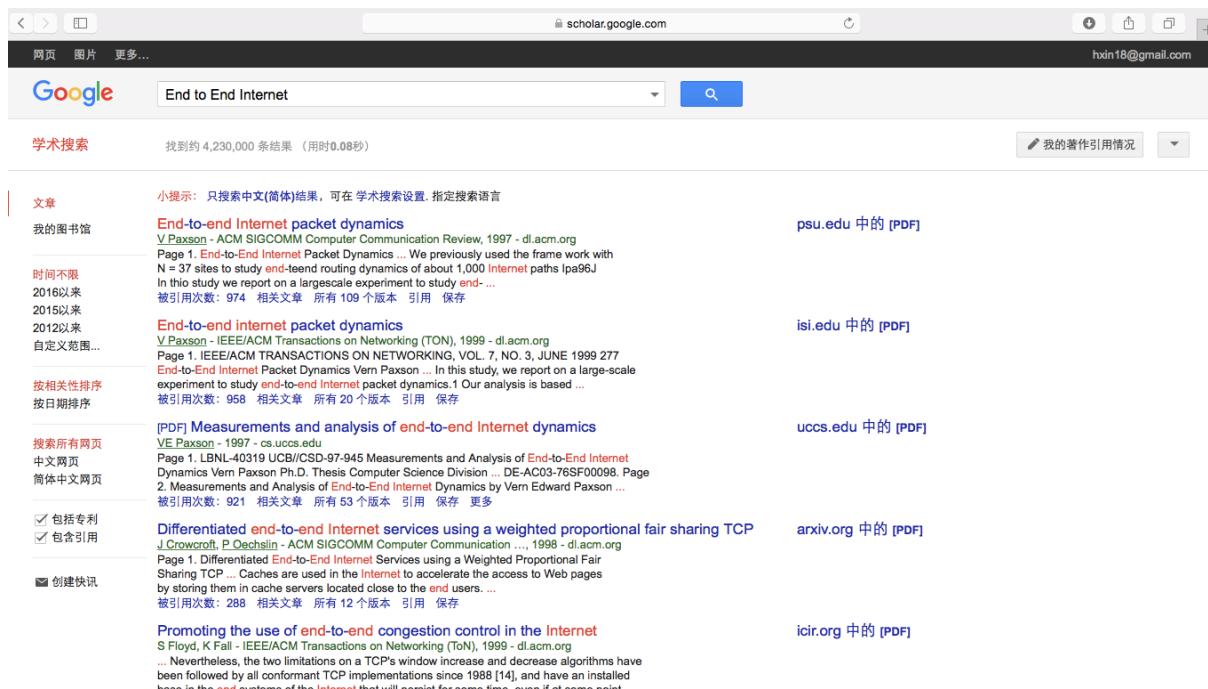
Table 5-2 Milestone papers for topic 26

Milestone Papers	Topic-Docs(%)
Random Early Detection Gateways for Congestion Avoidance	0.007842
Congestion Avoidance and Control	0.006743
Promoting the Use of End-to-End Congestion Control in the Internet	0.005095
Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications	0.004040
Equation-Based Congestion Control for Unicast Applications	0.003885
Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers	0.003739
End-to-End Internet Packet Dynamics	0.003377
A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing	0.003180
A Case for End System Multicast	0.003168
Rate Control for Communication Networks:Shadow Prices, Proportional Fairness and Stability	0.003118



The screenshot shows a Google Scholar search interface. The search term 'Congestion Avoidance Control' is entered in the search bar. The results page displays approximately 110,000 results found in 0.14 seconds. The results are categorized by document type (文章) and source (如 berkeley.edu, columbia.edu, academia.edu). Several results are highlighted with blue links to full-text PDFs or specific academic websites.

图 5–6 Google Scholar “Congestion Avoidance Control” 搜索结果
Fig 5–6 Search result of “Congestion Avoidance Control” on Google Scholar



The screenshot shows a Google Scholar search interface. The search term 'End to End Internet' is entered in the search bar. The results page displays approximately 4,230,000 results found in 0.08 seconds. The results are categorized by document type (文章) and source (如 psu.edu, isi.edu, uccs.edu, arxiv.org, icir.org). Several results are highlighted with blue links to full-text PDFs or specific academic websites.

图 5–7 Google Scholar “End to End Internet” 搜索结果
Fig 5–7 Search result of “End to End Internet” on Google Scholar

5.4.4 话题时间和热度展示

对于一个话题，话题的热门时间和热门程度是衡量话题兴盛时间和兴盛程度的重要指标。顾名思义，在生活中，话题热度代表着人们谈论某话题的可能性。当一个话题的热度越高，人们谈论它的概率就越大。在学术文献中，学术话题的热度是学者在撰写论文时比较频繁使用的主题。一定程度上，话题热度可以反映一个领域某话题的发展程度，当一个话题拥有很高的热度时，该话题往往处于高度发展的时刻。在这里，我们定义话题的热度为在一定时间内，文章选择主题的概率。模型中的文档—话题分布为我们参考的指标。

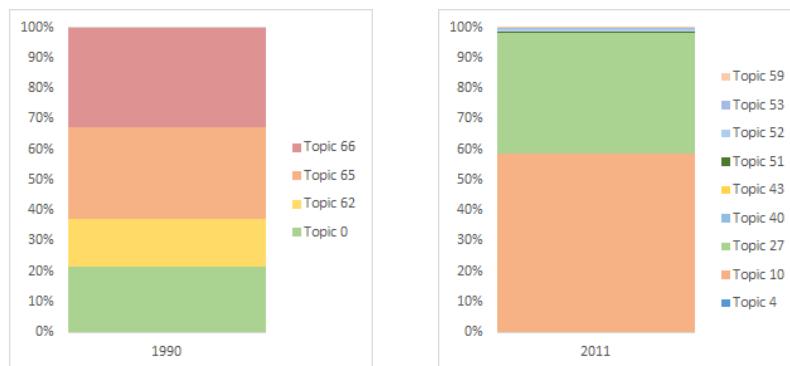


图 5-8 话题热度样例

Fig 5-8 Example of topic popularity

图5-8为AAN数据集中1990年前和2011年话题的热度的分布图。通过图中，我们可以发现在1990年前比较热门的主题有四个，分别是主题0，主题62，主题65和主题66。这四个主题的词云分别由图5-9(a)至图5-9(d)表示。从词云我们可以知道，在1990年前，自然语言处理领域中比较热门的话题包括“Nature Language System”，“Speech Recognition”，“Machine Translation”，和“Gramma Parsing”。在2011年，比较热门的主题有两个，分别是主题10和主题27。这两个主题的词云分别由图5-9(e)和图5-9(f)表示。从词云我们可以知道，在2011年，自然语言处理领域中比较热门的话题包括“Information Extraction”，和“Statistical Machine Translation”。我们注意到，1990年和2011年中“Machine Translation”均为重要的主题，但对比主题下的具体词语，我们发现1990的主题65的范围比较宽泛，并不集中于机器翻译的某一特定形式，而2011年主题27比较集中于统计机器翻译的范畴。由[32]中我们可以知道，1990年是基于规则的机器翻译发展成熟和统计机器翻译被提出并开始发展的阶段，因此机器翻译领域呈现的是比较多元的话题，而到了2011年，机器翻译领域的主要方法是统计机器翻译，而基于规则的机器翻译由于其局限性逐渐被淘汰，因此机器翻译的主要话题为统计机器翻译。在某种程度上，主题10和主题27在内容上的差异也反映了机器翻译领域主题的更迭。

对于一个话题，随着领域的不断发展，该主题会产生，发展，并在一定的时间到达热度的顶峰，并最终随着时间的发展而消亡。话题热度在时间的分布则是该主题的热度在时间上的分布。在本模型中，话题热度的参考指标是模型的话题-时间分布。为了体现不同话题热度的不同，展示话题随时间分布的时候将分布的概率值与该话题的总热度的乘积。

图5-10展示的是数据集中HEP-PH中关于量子物理，弦理论话题随时间的变化，由于物理领域

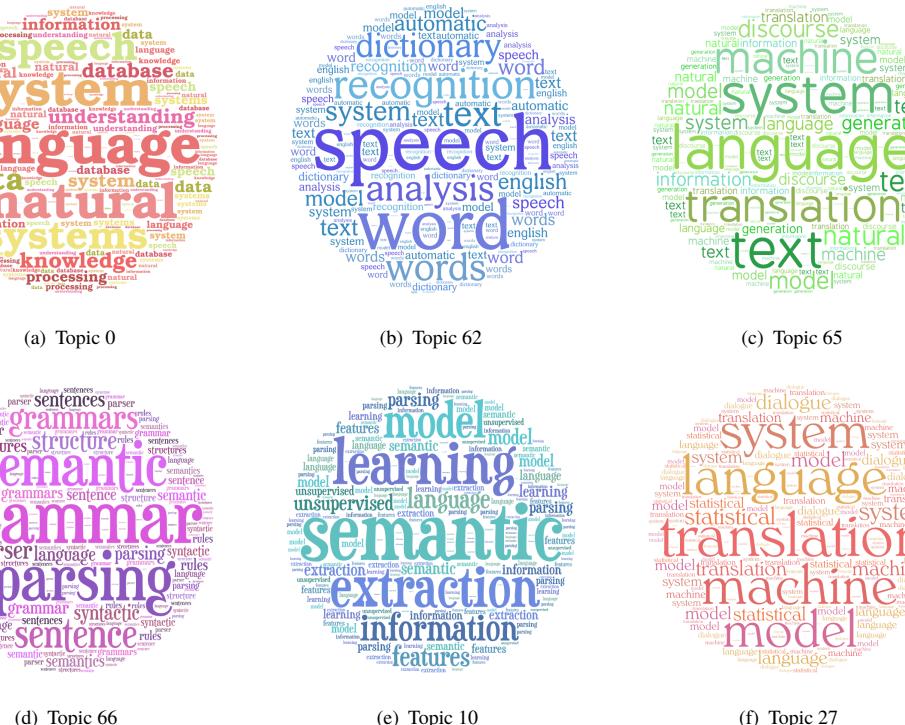


图 5-9 AAN 数据集话题样例
Fig 5-9 Example of topics in AAN Dataset

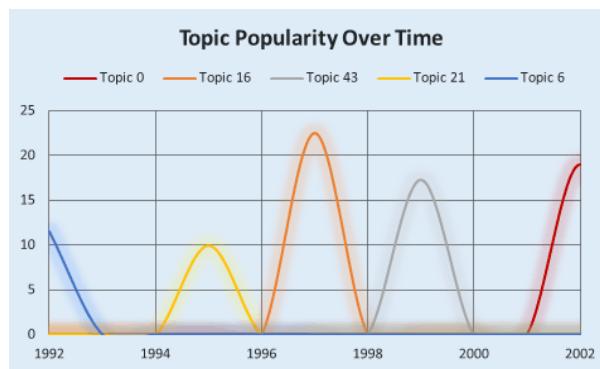


图 5-10 HEP-PH 话题热度随时间变化
Fig 5-10 Topic popularity over time in HEP-PH

已经高度成熟，所选的五个主题在词的分布上仅有细微区别，但从图里我们依旧可以观察出主题更迭的趋势。在这个领域中，主题 6，主题 21，主题 16，主题 43 和主题 0 从先到后达到了顶峰，然后逐步走向了消亡。

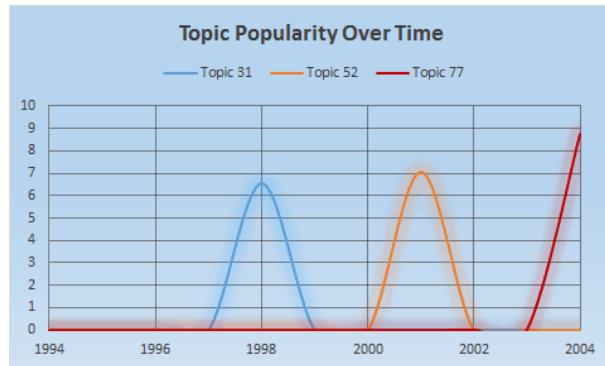


图 5-11 计算机网络话题热度随时间变化
Fig 5-11 Topic popularity of computer network over time

图5-11展示的是数据集中 *Citeseer* 中关于计算机网络话题随时间的变化，在这个领域中，热门话题发生了三次更迭，主题 31，主题 52 和主题 77 从先到后达到了顶峰，然后逐步走向了消亡。这三个主题的词云由图5-12(b)至5-12(d)所示，由词云可以看出，计算机网络领域的话题内容也发生了变化。

5.4.5 话题演变趋势展示

话题演变是某一个话题转变为另一个话题的过程，在这个过程的发生伴随着两个过程，话题的内容变化以及话题的热度变化。话题演变趋势的挖掘实际上是建立新旧话题之间联系的这一过程。在我们模型中，话题间的关系是通过文档相互引用的关系构建的，构建的途径为主题-社区-文档-主题，通过这一路径我们可以构成的主题-主题分布可以体现出对于给定一主题，该主题来源于另一主题的概率。在判断主题间演变强度时，主题-主题分布为我们所参考的指标，其表达式由公式4-8所示。

在5.4.4和中，我们已经初步发现 *AAN* 数据集中的机器翻译领域中话题发生的变化，在本节中，我们以 *AAN* 数据集中话题机器翻译为例，对话题演变趋势进行展示。

图5-13是我们在 *AAN* 数据集中抽取的主题中选取与机器翻译相关的四个主题。主题 65，主题 28，主题 63，主题 27 的时间点分别为 1990 年，2001 年，2006 年，2011 年。由 [32] 可知，机器翻译从 1990 年至今经历了基于文本的机器翻译的衰落和统计机器翻译的兴盛，而从图5-13中，我们发现，在 1990 年，机器翻译的范围比较宽泛，并没有深入进某一特定领域之中。进入 21 世纪，在主题 28 中我们发现了“statistical”和“corpus based”两个关键词，说明了在 2001 年，机器翻译主要分为统计机器翻译和基于语料库对机器翻译两大类，这两大类占据相近的比重。在 2006 年和 2011 年对应的主题中，我们发现“corpus”一词已经淡出话题-单词分布，这说明了这段时间内，机器翻译领域中占据主导的是统计机器翻译。

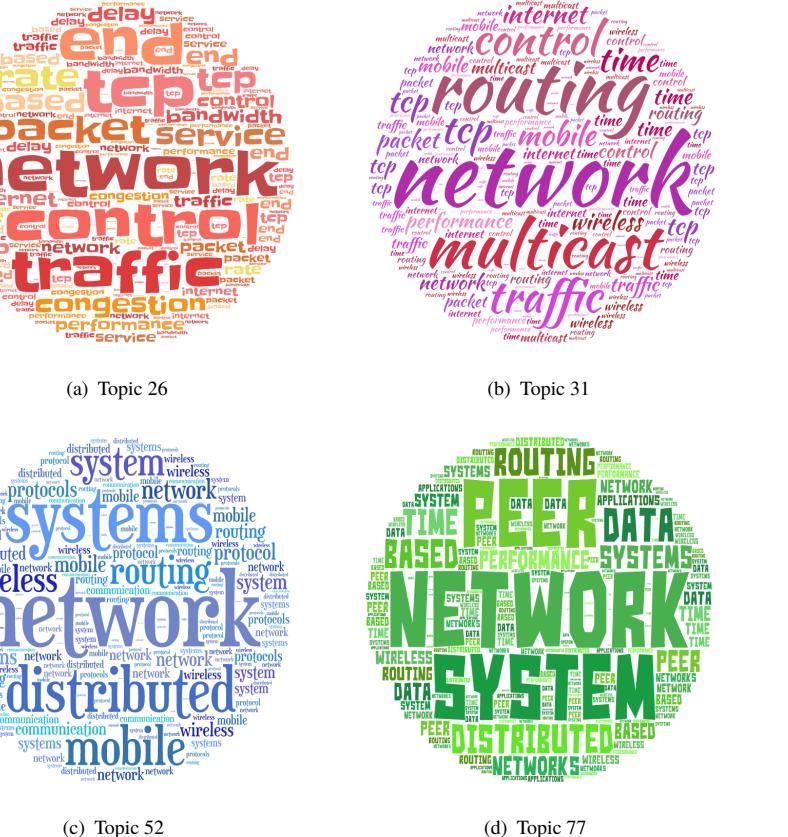


图 5-12 Citeseer 数据集话题样例
Fig 5-12 Example of topics in Citeseer Dataset

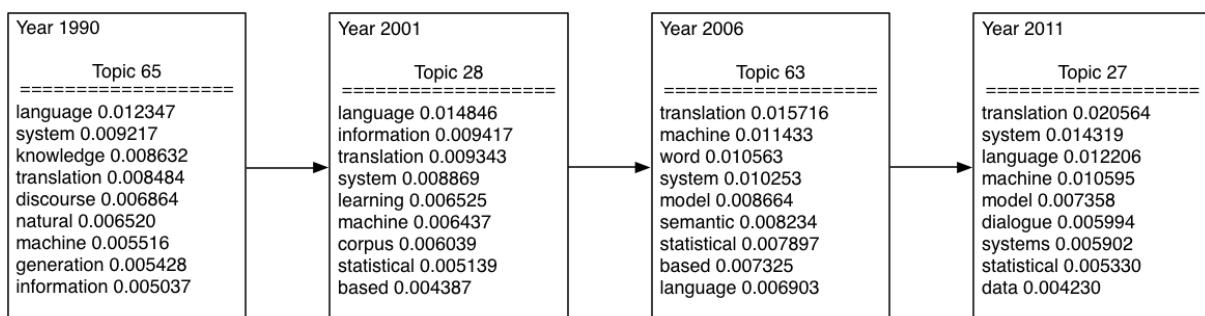


图 5-13 “机器翻译”话题演变趋势
Fig 5-13 Evolution pattern of “machine translation”

5.4.6 话题地图展示

学术话题演变的最终目标是展示出一个领域各话题的发展情况和演变过程，从5.4.2节到5.4.5节，我们分别研究了话题内容，话题下代表文章的推荐，话题热度及其变化趋势，话题的演变规则。将以上内容相结合，我们可以构建出一个领域的话题地图。

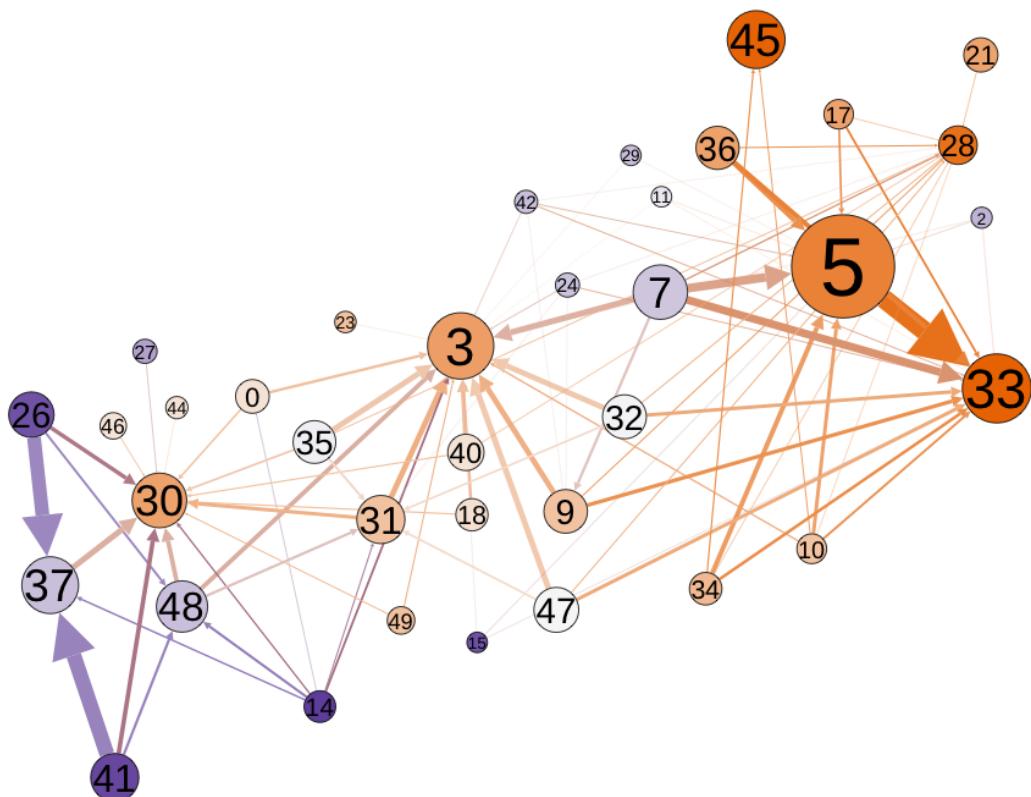


图 5-14 HEP-PH 主题地图

Fig 5-14 HEP-PH Topic Map

图5-14为 *HEP-PH* 数据集的主题地图。图中每个点代表一个主题，节点的大小是代表着主题的热门程度，主题的颜色则代表了主题的新旧。橙色代表旧主题，颜色越深则代表时间越早；紫色代表新主题，颜色越深则代表时间越后。主题间的箭头为主题之间的引用演变边。在图中，我们剔除了游离在聚类外的几个重要度较低的主题。主题边由新主题指向旧主题，边的宽度代表着主题间引用强度的大小。由于 *HEP-PH* 的主题集中在高能物理领域中，且主题间的相关度较高，因此，主题在地图里形成一个相互联系的整体，主题之间的相互联系也较为密集。这与 *HEP-PH* 训练集中文章间的引用网络极高的密集程度息息相关。

图5-15为 *AAN* 数据集的主题地图。图中每个点代表一个主题，节点的大小是代表着主题的热门程度，主题的颜色则代表了主题的新旧。红色代表旧主题，颜色越深则代表时间越早；蓝色代表新主题，颜色越深则代表时间越后。主题间的箭头为主题之间的引用演变边。在图中，我们剔除了游离在聚类外的几个重要度较低的主题。主题边由新主题指向旧主题，边的宽度代表着主题间引用强

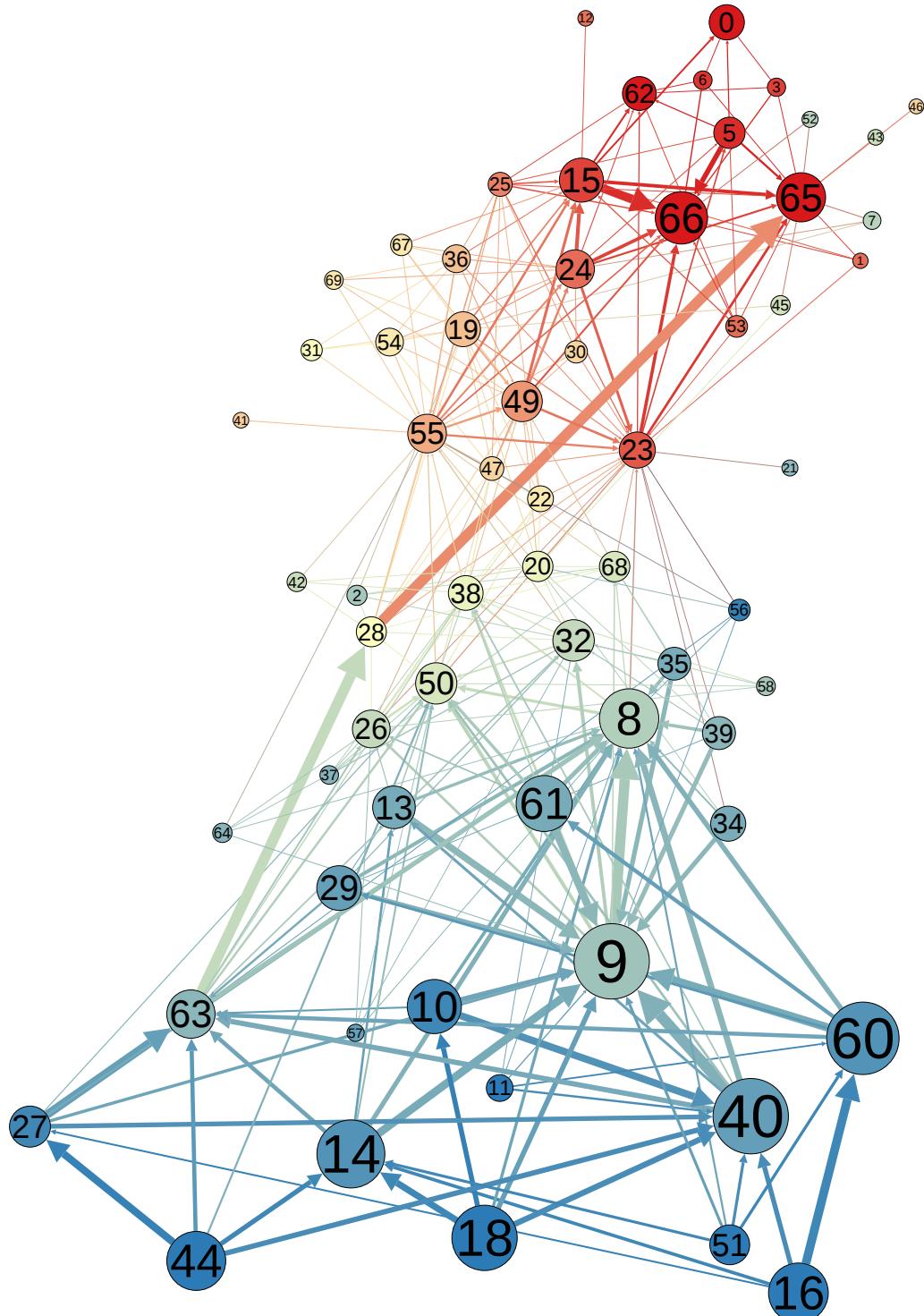


图 5-15 AAN 话题地图

Fig 5-15 AAN Topic Map

度的大小。由于 AAN 数据集的主题集中在自然语言处理领域中，且主题间的相关度较高，因此，主题在地图里形成一个相互联系的整体，主题之间的相互联系也较为密集，但不同的主题间的引用边的宽度有着明显的区别。与此同时，我们可以观察出机器翻译的发展趋势：主题 44 来源于主题 27，最终沿着主题 63，主题 28 的路径，最终追溯到主题 65。



图 5-16 Citeseer 话题地图
Fig 5-16 Citeseer Topic Map

图5-16为 Citeseer 数据集的主题地图。图中每个点代表一个主题，节点的大小是代表着主题的热门程度，主题的颜色则代表了主题的新旧。黄色代表旧主题，颜色越深则代表时间越早；绿色代表新主题，颜色越深则代表时间越后。主题间的箭头为主题之间的引用演变边。在图中，我们剔除了游离在聚类外的几个重要度较低的主题。主题边由新主题指向旧主题，边的宽度代表着主题间引用强度的大小。由于 Citeseer 的主题集中在计算机中，因此主题之间联系比较频繁，但与 AAN 数据

集和 *HEP-PH* 相比, *Citeseer* 数据集中的主题可以分为以主题 53, 主题 19 和主题 13 为中心的三个部分。其中主题 53 的关键词包括 “*performance*”, “*network*”, “*memory*”, “*data*”, “*parallel*” , 偏向计算机系统和网络方向; 主题 19 的关键词包括 “*data*”, “*algorithm*”, “*logic*”, “*model*”, “*query*”, “*based*” , 偏向数据科学方向; 主题 19 的关键词包括 “*algorithm*”, “*learning*”, “*linear*”, “*model*” , 较为理论。而在数据科学的研究中, 我们常常应用到偏理论的算法, 这一特点也从地图中主题 19 和主题 13 联系较为密切体现出来。

5.5 实验结果总结

综合5.4节所得的结果, 我们认为所设计的联合主题模型在提取文本主题, 寻找主题相关文章, 探究主题时间和热度变化, 揭示主题间演变关系上有令人满意的表现, 并成功构建反映领域发展趋势的主题地图。通过模型所产生的参数, 我们可以准确地把握学术数据中话题演变信息, 掌握领域发展的规律。

第六章 结语

6.1 课题总结

本课题设计并实现了一个基于 LDA 模型的联合主题模型，并将该模型应用在探究学术话题演变上。模型的设计综合考虑了学术论文的文本信息，引用信息和时间信息，并在模型的生成过程中建立了各个参数之间的联系方式。通过参数的相互组合，模型完成了提取文本主题，寻找主题相关文章，探究主题时间和热度变化，揭示主题间演变关系，并最终构建主题地图的任务。与此同时，模型在真实学术数据集中的应用上有不错的表现。

简言之，本课题有以下贡献：

1. 设计新的联合主题模型，模型将学术数据中的文本信息，引用信息和时间信息有机结合。
2. 完成模型相关代码框架的编写和设计。
3. 通过模型训练所得到的参数，回答了学术话题演变的基本问题。
4. 模型成果应用在三个学术大数据数据集中，完成学术主题提取，关键文章查找，话题热度变化挖掘，演变趋势的发现，并取得令人满意的结果。
5. 在研究学术话题演变的过程中，构建了学术话题地图，更加直观反映领域的构成和演变趋势。

6.2 未来研究方向

学术大数据话题演变挖掘是学术大数据分析的重要组成部分，它有着很高的价值和应用空间。因此，未来将会有越来越多的工作促进学术话题演变挖掘的发展和完善。在总结调研课题所阅读的文献和完成课题所使用的方法和模型后，我们认为，学术话题演变的分析可以在以下方向继续挖掘和发展。

1. 主题模型是一个非监督的模型，这也使得模型所得结果的评判是一个较为主观的过程，模型训练结果效果的好坏主要依靠人的直觉判断。在研究话题演变这一课题中，这个问题同样存在。在已有工作中并没有针对学术话题演变设立的评价指标，这也使得话题演变的评价指标设计成为了一个意义重大的未来方向
2. LDA 模型是遵循着词袋假设的非监督模型，尽管这一模型在主题提取上具有较好的性能，但词袋模型忽略词序的特点是该模型遭受诟病的一点。因此，如何应用其他主题提取方法提取主题并应用在学习话题演变上也值得我们关注。
3. 目前的学术大数据中，数据资料并不完善。如何在部分文章存在信息缺失(如文本信息缺失)的情况下仍然保证主题挖掘的质量，并展示研究主题演变的趋势也是一个应用广泛的方向。

附录 A 模型推断与参数估计

这个附录将说明第四章中所提出模型的推断过程。在4.3.5节中我们定义了变分推断所构造的分布 q :

$$q(\theta, z, z', c | \gamma, \phi, \lambda, \sigma) = q(\theta | \gamma) \prod_{n=1}^{N_d} q(z_n | \phi_n) \prod_{l=1}^{L_d} q(z'_l | \lambda_l) \prod_{l=1}^{L_d} q(c_l | \sigma_l) \quad (\text{A-1})$$

与 [6] 相同，我们首先使用琴声不等式写出模型对数似然值的下界。出于简化公式的目的，我们在写法上忽略变分变量。

$$\begin{aligned} \ln P(w, l, t | \alpha, \beta, \psi, \Omega, \eta) &= \ln \int \sum_z \sum_{z'} \sum_c P(w, l, t, \theta, z, z', c | \alpha, \beta, \psi, \Omega, \eta) d\theta \\ &= \ln \int \sum_z \sum_{z'} \sum_c \frac{P(w, l, t, \theta, z, z', c) q(\theta, z, z', c)}{q(\theta, z, z', c)} d\theta \\ &\geq \int \sum_z \sum_{z'} \sum_c q(\theta, z, z', c) \ln P(w, l, t, \theta, z, z', c | \alpha, \beta, \psi, \Omega, \eta) d\theta \\ &\quad - \int \sum_z \sum_{z'} \sum_c q(\theta, z, z', c) \ln q(\theta, z, z', c) \theta \\ &= E_q[\ln P(w, l, t, \theta, z, z', c | \alpha, \beta, \psi, \Omega, \eta)] - E_q[\ln q(\theta, z, z', c)] \quad (\text{A-2}) \end{aligned}$$

所得的下界与公式4-4a中的 $L(q)$ 为同一概念。我们将公式A-2进行展开，得到公式A-3。

$$\begin{aligned} L(\gamma, \phi, \lambda, \sigma; \alpha, \beta, \psi, \Omega, \eta) &= E_q[\ln P(\theta | \alpha)] + E_q[\ln P(\mathbf{z} | \theta)] + E_q[\ln P(\mathbf{w} | \mathbf{z}, \beta)] + E_q[\ln P(\mathbf{t} | \mathbf{z}, \psi)] \\ &\quad + E_q[\ln P(\mathbf{z}' | \theta)] + E_q[\ln P(\mathbf{c} | \mathbf{z}', \eta)] + E_q[\ln P(\mathbf{l} | \mathbf{c}, \Omega)] \\ &\quad - E_q[\ln q(\theta | \gamma)] - E_q[\ln q(\mathbf{z} | \phi)] \\ &\quad - E_q[\ln q(\mathbf{z}' | \lambda)] - E_q[\ln q(\mathbf{c} | \sigma)] \quad (\text{A-3}) \end{aligned}$$

将公式A-3逐一展开，我们可以得到：

$$\begin{aligned}
 L(\gamma, \phi, \lambda, \sigma; \alpha, \beta, \psi, \Omega, \eta) &= \ln(\Gamma(\sum_{i=1}^K \alpha_i)) + \sum_{i=1}^K \log(\Gamma(\alpha_i)) + \sum_{i=1}^K (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \\
 &\quad + \sum_{n=1}^N \sum_{i=1}^K \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \\
 &\quad + \sum_{n=1}^N \sum_{i=1}^K \sum_{j=1}^V \phi_{ni} w_n^j \ln(\beta_{ij}) \\
 &\quad + \sum_{n=1}^N \sum_{i=1}^K \phi_{ni} w_d^t \ln(\psi_{it}) \\
 &\quad + \sum_{l=1}^L \sum_{i=1}^K \lambda_{li}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \\
 &\quad + \sum_{l=1}^L \sum_{i=1}^K \sum_{j=1}^C \sigma_{lj} \lambda_{li} \ln(\eta_{ij}) \\
 &\quad + \sum_{l=1}^L \sum_{i=1}^C \sum_{j=1}^D c_l^j \sigma_{li} \ln(\Omega_{ij}) \\
 &\quad - \ln(\Gamma(\sum_{i=1}^K \gamma_i)) + \sum_{i=1}^K \log(\Gamma(\gamma_i)) - \sum_{i=1}^K (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \\
 &\quad - \sum_{n=1}^N \sum_{i=1}^K \phi_{ni} \ln(\phi_{ni}) \\
 &\quad - \sum_{l=1}^L \sum_{i=1}^K \lambda_{li} \ln(\lambda_{li}) \\
 &\quad - \sum_{l=1}^L \sum_{i=1}^C \sigma_{li} \ln(\sigma_{li})
 \end{aligned} \tag{A-4}$$

A.1 变分推断

在本节中，我们将求解最大化公式A-3所对应的变分变量 ϕ , γ , λ 和 σ 。由于这四个变量均存在行归一化的性质，变分推断的本质实际上是一个受约束的最优化问题。因此，我们采用拉格朗日数乘法进行最大化求解：

- 对于 ϕ_{ni} , 在所需最大化的公式由A-13所示

$$L_{[\phi_{ni}]} = \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) + \phi_{ni} \ln \beta_{in_i} + \phi_{ni} \ln \psi_{it_d} + \lambda (\sum_{j=i}^K \phi_{nj} - 1) \tag{A-5}$$

在这里, 对 ϕ_{ni} 进行求导, 我们可以得到,

$$\frac{\partial L_{[\phi_{ni}]}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j) + \ln \beta_{in_i} \psi_{it_d} - \ln \phi_{ni} - 1 + \lambda \tag{A-6}$$

因此，我们有

$$\phi_{ni} \propto \beta_{in_i} \psi_{it_d} \exp(\Psi(\gamma_i)) \quad (\text{A-7})$$

类似的对于同属于变分多项分布多变量 λ 和 σ ，我们可以得到：

$$\lambda_{li} \propto \exp(\Psi(\gamma_i) + \sum_{c=1}^C \sigma_{lc} \ln(\eta_{ic})) \quad (\text{A-8})$$

$$\sigma_{lj} \propto \Omega_{jl_l} \exp\left(\sum_{k=1}^K \lambda_{lj} \ln(\eta_{kj})\right) \quad (\text{A-9})$$

- 对于变分狄利克雷变量 γ 的推断，其需要最大化的公式由A-14所示

$$\begin{aligned} L_{[\gamma]} &= \sum_{i=1}^K (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) + \sum_{n=1}^N \sum_{i=1}^K \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \\ &\quad + \sum_{l=1}^L \sum_{i=1}^K \lambda_{li} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \\ &\quad - \ln(\Gamma(\sum_{i=1}^K \gamma_i)) = \sum_{i=1}^K \log(\Gamma(\gamma_i)) + \sum_{i=1}^K (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \end{aligned} \quad (\text{A-10})$$

在这里，对 γ_i 进行求导，我们可以得到

$$\begin{aligned} \frac{\partial L_{[\gamma_i]}}{\partial \gamma_i} &= \Psi'(\gamma_i)(\alpha_i + \sum_{n=1}^N \phi_{ni} + \sum_{l=1}^L \lambda_{li} - \gamma_i) \\ &\quad - \Psi'\left(\sum_{j=1}^K \gamma_j\right) \sum_{j=1}^K (\alpha_j + \sum_{n=1}^N \phi_{nj} + \sum_{l=1}^L \lambda_{lj} - \gamma_j) \end{aligned} \quad (\text{A-11})$$

将 A-11取零，我们可得

$$\gamma_i = \alpha + \sum_{n=1}^N \phi_{ni} + \sum_{l=1}^L \lambda_{li} \quad (\text{A-12})$$

A.2 参数推断

在本节中，我们将求解最大化公式A-3所对应的模型变量 α , β , ψ , Ω 和 η 。

- 对于 α 的推断，我们依照 [6] 中的 Newton-Raphson 算法进行更新。对于其他四个变量，更新的方式和A.1中一致，使用拉格朗日数乘法。
- 对于 β_{ij} ，在所需最大化的公式由A-13所示

$$L_{[\beta_{ij}]} = \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{i=1}^K \sum_{j=1}^V w_{dn}^j \phi_{dn} \ln(\beta_{ij}) + \sum_{i=1}^K \lambda_i (\sum_{j=1}^V \beta_{ij} - 1) \quad (\text{A-13})$$

在这里，对 β_{ij} 进行求导，我们可以得到，

$$\frac{\partial L_{[\beta_{ij}]}}{\partial \beta_{ij}} = \frac{\sum_{d=1}^D \sum_{n=1}^{N_d} w_{dn}^j \phi_{dn} i}{\beta_{ij}} + \lambda_i \quad (\text{A-14})$$

将公式A-14取零，我们有：

$$\beta_{kv} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} w_{dn}^v \phi_{dn} k \quad (\text{A-15})$$

类似地，我们也可以得到其他三个变量的迭代公式：

$$\psi_{kt} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} w_d^t \phi_{dn} k \quad (\text{A-16})$$

$$\eta_{kc} \propto \sum_{d=1}^D \sum_{l=1}^{L_d} \sigma_{dlc} \lambda_{dlk} \quad (\text{A-17})$$

$$\Omega_{cp} \propto \sum_{d=1}^D \sum_{l=1}^{L_d} c_{dl}^p \sigma_{dlc} \quad (\text{A-18})$$

参考文献

- [1] Amr Ahmed and Eric P Xing. “Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process: with Applications to Evolutionary Clustering.” In: *SDM*. SIAM. 2008, pp. 219–230.
- [2] Edoardo M Airoldi et al. “Mixed membership stochastic blockmodels”. In: *Journal of Machine Learning Research* 9.Sep (2008), pp. 1981–2014.
- [3] David Blei and John Lafferty. “Correlated topic models”. In: *Advances in neural information processing systems* 18 (2006), p. 147.
- [4] David M Blei and John D Lafferty. “Dynamic topic models”. In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 113–120.
- [5] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Advances in neural information processing systems*. 2001, pp. 601–608.
- [6] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.
- [7] Sidney Coleman, John Preskill, and Frank Wilczek. “Quantum hair on black holes”. In: *Nuclear Physics B* 378.1 (1992), pp. 175–246.
- [8] Scott Deerwester et al. “Indexing by latent semantic analysis”. In: *Journal of the American society for information science* 41.6 (1990), p. 391.
- [9] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the royal statistical society. Series B (methodological)* (1977), pp. 1–38.
- [10] Elena Erosheva, Stephen Fienberg, and John Lafferty. “Mixed-membership models of scientific publications”. In: *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), pp. 5220–5227.
- [11] William A Gale, Kenneth W Church, and David Yarowsky. “One sense per discourse”. In: *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics. 1992, pp. 233–237.
- [12] DMBTL Griffiths and MIJJB Tenenbaum. “Hierarchical topic models and the nested chinese restaurant process”. In: *Advances in neural information processing systems* 16 (2004), p. 17.
- [13] Thomas L Griffiths and Mark Steyvers. “Finding scientific topics”. In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.
- [14] Qi He et al. “Detecting topic evolution in scientific literature: how can citations help?” In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 957–966.

- [15] Thomas Hofmann. "Probabilistic latent semantic analysis". In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1999, pp. 289–296.
- [16] Peder Olesen Larsen and Markus Von Ins. "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index". In: *Scientometrics* 84.3 (2010), pp. 575–603.
- [17] Shawn Martin et al. "OpenOrd: an open-source toolbox for large graph layout". In: *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics. 2011, pp. 786806–786806.
- [18] Christof Monz. "Statistical machine translation with local language models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, pp. 869–879.
- [19] Ramesh Nallapati and William W Cohen. "Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs." In: *ICWSM*. 2008.
- [20] Ramesh M Nallapati et al. "Joint latent topic models for text and citations". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 542–550.
- [21] Ramesh M Nallapati et al. "Multiscale topic tomography". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2007, pp. 520–529.
- [22] Kamal Nigam et al. "Text classification from labeled and unlabeled documents using EM". In: *Machine learning* 39.2-3 (2000), pp. 103–134.
- [23] Dragomir R Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. "The ACL anthology network corpus". In: *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Association for Computational Linguistics. 2009, pp. 54–61.
- [24] Svitlana Volkova et al. "Lightly Supervised Learning of Procedural Dialog Systems." In: *ACL (1)*. Citeseer. 2013, pp. 1669–1679.
- [25] Chong Wang, David Blei, and David Heckerman. "Continuous time dynamic topic models". In: *arXiv preprint arXiv:1206.3298* (2012).
- [26] Xiaolong Wang, Chengxiang Zhai, and Dan Roth. "Understanding evolution of research themes: a probabilistic generative model for citations". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 1115–1123.
- [27] Xuerui Wang and Andrew McCallum. "Topics over time: a non-Markov continuous-time model of topical trends". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2006, pp. 424–433.
- [28] Jonathan Stuart Ward and Adam Barker. "Undefined by data: a survey of big data definitions". In: *arXiv preprint arXiv:1309.5821* (2013).

- [29] Kyle Williams et al. “Scholarly big data information extraction and integration in the CiteSeer χ digital library”. In: *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*. IEEE. 2014, pp. 68–73.
- [30] Jaewon Yang and Jure Leskovec. “Community-affiliation graph model for overlapping network community detection”. In: *2012 IEEE 12th International Conference on Data Mining*. IEEE. 2012, pp. 1170–1175.
- [31] Jaewon Yang, Julian McAuley, and Jure Leskovec. “Community detection in networks with node attributes”. In: *2013 IEEE 13th International Conference on Data Mining*. IEEE. 2013, pp. 1151–1156.
- [32] 刘群. “机器翻译研究新进展”. In: *当代语言学* 2 (2009), pp. 147–158.

致 谢

首先我要向我的导师王新兵教授表达最诚挚的谢意的敬意。本论文的顺利完成离不开王老师的悉心指导。在选择本课题之间，我对课题所涉及的关于主题模型，数据挖掘等知识都了解甚浅，在王老师的鼓励和支持下，我勇敢地尝试了这一个较为陌生的课题，挑战了自己，也在王老师所带领的 Acemap 组里接受到完整而系统的训练，对此我感到非常幸运。

课题的提出和探索是一个枯燥艰辛的过程，如何在一个发展较为成熟的领域添加自己的理解并有所提高需要克服许多困难。在完成课题的过程中，我深深知道，如果没有组里许多同学的帮助和指导，我所得到的结果将会大打折扣。在这里，我要尤其感谢沈嘉明组长，何俊贤同学黄颖同学及其理论组的其他同学，感谢同学们在这期间的陪伴，尤其在模型的建立，推导上，和代码的编写上给了我莫大的帮助和支持。除此之外，我还要感谢大数据组的马松君学长和 Spark 组的陈戈学长。在我进入实验室的两年来所取得的进步离不开他们的指导和督促。

我也要感谢巴黎高科卓工程师学院的各位同学和我本科期间遇到的导师。正是因为他们，我才拥有一个充实丰富的大学四年生活。

我要感谢我的父母，他们是我最坚强的后盾，在我遇到困难的时候，无条件地为我最大的支持和帮助。让我每时每刻都能感受到温暖。

最后，我要感谢交通大学四年来对我的培养，她教会了我知识，交给我责任，带给我成长。她的精神鼓舞着我，让我向未来坚定地走下去。

Topic Evolution in Scholarly Big-data

With the development of computer science, the human society has stepped into digital age. In this era, data and documents are stored and transferred in electronic form. These digital materials have contributed to the formation of big-data era. In 21st century, with the advancement of Internet and social network, more and more human activities appear online and produce a great number of data, accelerating the development of big-data era.

The big-data era also has its impact on scholar field. Every year, the scholars and research groups continue to produce large amount of scholarly materials, which include papers, text books, patents, etc. With this trend, there is a great number of available data online. For example, the Citeseer dataset contains over 7 million papers, in addition, the amount of available papers on Google Scholar are believed to be even more. The production and accumulation of scholarly data have a positive impact on the society as it increases the exchanges of knowledges, promotes the communications among researchers and thus makes research activities more effective. Under this background, there is a growing tendency that scholarly big-data analysis has become an important topic in the data-mining field.

The studies of scholarly big-data contain a wide range of research directions, including understanding how the scientific topics evolve, appear or extinct; how to judge the importance of scientific literatures; how to uncover the relations among papers, research groups and scientific fields, etc. In this paper, we mainly focus on understanding the topic evolution in scholarly big-data. The objective of mining topic evolution patterns in scholarly big-data is to extract keywords from the data to represent the academic topics and to reveal the evolution patterns among them. The evolution patterns not only include the changes of topic content over time, but also involve the whole procedure of how topics appear, develop and emerge. Further more, it is important to understand how old topics turn into new ones and what are the dependency among topics in different periods. The final goal of studying topic evolution in scholarly big-data is to show the whole picture of academic fields.

Understanding how topics develop in a scientific area is of great importance with a wide range of applications. In the era of scholarly big-data, with the number of available papers on the Internet grows day by day, failing to acknowledge how a field forms and develops may cause researchers, who start a new research, struggle in finding the relevant papers. However, if the evolution patterns of topics are well understood, it is easier for those researchers to find out the related topics to their researches. They can not only understand what are the topics about, but also how they develop and what are the milestone papers. In this way, the researchers are able to get familiar with the new scientific fields in a much quicker way.

Topic evolution has been extensively studied before and related works include [1, 4, 21, 25, 27]. However, these works mainly focus on the change of topic content over time (eg. what are the keywords of topics in different periods) or how to determine the periods in which a topic is popular. However, these works rarely

try to uncover the evolution patterns among topics. From our points of view, the evolution patterns among topics, especially the relations between the old topics and new topics, reflect the structure and development of the scientific field in a more direct and vivid way. Another limit for using these models directly on the analysis of academic big-data is that these models merely focus on the textual information in the data, while the citation network also plays an important role in scholarly data. On one hand, the citation information contains an implicit temporal information as, in the reality, the papers can only cite papers that are published earlier; on the other hand, when a researcher write a paper, it is of high possibility that he cites the paper belonging to the similar research fields. Papers citing older papers with similar topics, to some extent, can be considered as a way for topics to evolve. Based on these concerns, [26] adopted the assumption of “bag of citations”, analogous to the “bag of words” assumption, and used topic model to generate the citation network of scientific literatures. Training the parameters in the model, the author built links between citation topics (clusters of documents). Besides, the author has also formalised the answers to the principal questions in studying the evolution of scientific themes:

1. How to present the topics in scientific literatures?
2. How to find milestone papers for scientific topics?
3. How to find the most dominant topics that are extensively studied?
4. When did the topics become popular and are they still attracting attention today?
5. How to uncover the dependency among topics?
6. How to identify the underlying evolution patterns among topics?

However, in the model proposed in [26], the author only took the citation network of papers into consideration, but failed to consider the textual information in the generation process of its model. In order to present the topic keywords, he just extracted the keywords from titles and abstracts of papers, this process may result in missing some important information(eg. terminology). Besides, due to the characteristics of citation data, the frequency of generating the old documents is much higher than the frequency of new papers, intuitively, the topic year discovered in [26] tends to be earlier than the actual time-stamp.

In this paper, we mainly focus on revealing the evolution patterns in academic data by answering the questions listed above. In order to accomplish that, we propose a novel joint topic model which combines the textual information, the reference and the temporal information of into the generation process of a document. The model is constructed based on the Latent Dirichlet Allocation(LDA)^[26] and the generation process of the textual information in a document follows the similar generation mechanism in LDA model. In the process of discovering network community of citation network, we adopt the similar mechanism with the model proposed in [26]. Besides, instead of mixing the concept of topics and network community together, we consider these two concepts differently and introduce a parameter to build a link between them. In the aspect of discovering temporal information of topic, we modify the generation process of every word in a document compared with the standard procedure in LDA model. We divide the time into several periods, as soon as a word is generated, we assign a period associated with the topic that have been assigned to the word. In this way, we are able to describe the periods in which a topic becomes popular via topic-period distribution. We derive the model by variational EM algorithm and implement the model with a C++ program. With training data, we can learn

the parameters involved in our model. Using these parameters, we construct a formalised solutions to the questions listed above.

In the experiment, we test our model on three different academic datasets which are respectively ACL Anthology Network(AAN) dataset, Arxiv HEP-PH dataset as well as a subset of Citeseer dataset containing 51208 documents in the field of computer science. We demonstrate the result of our model in this paper and prove the effectiveness of our model. In the aspect of topic extraction, we find that the topics that we extract are of relatively high quality. Topics are correctly assigned to documents and that topic keywords are clear and belong to a concrete scientific field. In the part of milestone papers discovering, we conduct two case studies and compared the retrieved results of Google Scholar, searching the topic keywords. The milestones papers returned by our model in the topic of *Statistical Machine Translation*, and the topic of *End to End Internet Congestion Control* are included in the top search results of Google Scholar. In the study of topic popularity over time, we plot the change of popularity over time of topics in two different fields respectively in the Arxiv HEP-PH dataset and Citeseer dataset. To show the results of discovering the evolution patterns, we dive into the topic *Machine Translation*, and find that the evolution patterns we observe are similar to a survey which describes how this field develop. Finally, we present the topic map of the three datasets, which show the whole picture of the fields in these three datasets.

To be concrete, the main contributions of this paper can be concluded in the following aspects.

- We propose a new joint topic model. The model successfully combines the textual information, reference information and publish year of a document. We derive the model with variational EM algorithm and implement with C++ programs.
- We formalise the answers to the questions associated with topic evolution in scholarly big-data using the parameters in the model.
- We applicate our model in three different scholarly datasets, the result demonstrates that our model have a performance of high quality in the aspects of topic extraction, milestone papers discovery and topic evolution patterns mining.
- We present topic maps for all three academic datasets, showing how scientific fields are constructed and how they develop.

To sum up, the proposition and implementation of novel joint topic model the formalisations of the solutions to the principal questions are intuitive and clear, and the results of application in real scholarly datasets are reletively of high quality. The overall performance of the whole framework is satisfying.