● 叶春蕾[1,2]，冷伏海[2]
(1. 北京城市学院　信息学部，北京　100094；2. 中国科学院　国家科学图书馆，北京　100190)

基于引文—主题概率模型的科技文献主题识别方法研究

摘　要： 海量的科技文献中蕴含着大量揭示学科内容的主题信息。文章提出了一种新的概率模型：引文—主题概率模型，该模型对文献中的关键词和引文进行联合建模以完成科技文献中的主题内容识别，在获得主题中关键词分布的同时也获得相关主题间的引文分布。实验表明，基于引文—主题模型识别的主题信息能为进一步的主题演化分析提供一定的分析基础。

关键词： 科技文献；主题识别；引文—主题模型；方法研究

Abstract: There is abundant topic information in mass scientific literatures which can reveal the content of the subject. This paper proposes a new probabilistic model, the citation-topic model, which jointly models the keywords and citations in scientific literatures to complete the topic identification. The keyword distribution among the topics as well as the citation distribution among the related topics are obtained from the model. The experiment shows that the topic information identified by the citation-topic model can provide the basis for further topic evolution analysis.

Keywords: S & T literature; topic identification; citation-topic model; method study

随着网络技术的发展，情报研究人员可以利用情报学或其他学科的相关研究方法对海量文献进行全方位的信息分析，挖掘文献中隐性知识，正确识别科技领域研究的主题内容，为动态跟踪科技领域技术主题的演化规律、正确识别技术发展的水平等级打下基础。本文以相关理论为依据，在对现有研究进行深入分析的基础上，提出一种基于引文—主题概率模型（Citation-Topic Model，CTM）的主题识别方法，以解决主题识别的完整性和准确性问题，为进一步的主题演化分析提供一定的分析基础。

1　LDA 的主题概率模型

2003 年，Blei 等人提出了 LDA（Latent Dirichlet Allocation）模型[1]。LDA 是一种 3 层贝叶斯概率模型，包含词、主题、文档 3 层结构。该模型可以很好地模拟文档的生成过程，建立在概率层次下的主题识别能准确地表达词的语义层次关系，而不需要额外的词表开销，能更精确地把握主题识别过程，并对主题分析以及主题预测有很好的效果。Blei 等提出 LDA 主题模型的主要目的是希望能在保持基本统计关系的大规模数据集中发现小规模的关键词，而这些关键词能有效地用于分类、挖掘、概要以及相关性计算等的研究中。

目前国内外研究者广泛使用 LDA 模型进行主题信息的识别。T. L. Griffiths 等使用 LDA 模型识别科技文献中的主题信息[2]，石晶等使用 LDA 模型对文档的主题词进行抽取，以挖掘文本的主题内涵[3]，王金龙等使用 LDA 模型抽取科技文献的主题，为进一步主题演化分析提供基础[4]。笔者前期也使用 LDA 模型对科技文献中的主题识别进行了实证研究，证实 LDA 具有较好的主题识别能力。

科技文献中所包含的信息不仅是词的集合，对于如何完整地识别文献中包含的主题内容来说，引文作为一个重要组成要素很自然地应该参与到文献的主题内容构成中。因此，将引文引入主题内容不仅能够丰富主题内容，更能全方位地展示主题间的关联性。Bolelli 等提出作者—主题模型进行主题识别，使用作者去识别和提高文档中表征主题特征的主题词权重[5-6]。王萍提出 Topic-Author 模型，该模型对文献的文本信息和作者信息进行建模，并提出了多维度文献知识挖掘方法[7]。J. Yookyung 等尝试在主题识别时将引文和文本相结合[8]，如果一个短语和某一主题相关，那么包含这个词的文档所构成的引文子图比其他仅包含随机选择的文档所构成的子图具有更高的密度，换句话说，引文仅仅用于量化文献之间的主题相似度。

引文参与主题识别的关键是如何将引文直接转换为表征主题内容的特征，从直观角度来看，当一篇文献 A 引用另外一篇文献 B 时，A 往往使用 B 的内容扩展自身主题信息，因此，B 中的主题应该对 A 中的主题有一定的影响。在对文献 A 进行主题识别时，如果不考虑这些影响，很可能会丢失一些和 A 相关的主题信息。因此，本文提出的引文—主题概率模型是一种多重概率模型，该模型在

文献集 $D(t)$ 中识别主题时，除考虑 $D(t)$ 中文献本身产生的关键词，还考虑 $D(t)$ 引用的文献集 $D'(t)$ 的主题特征。本文尝试在以科技文献中的关键词为核心构建的贝叶斯框架中直接引入引文，以文献的关键词和引文共同建立引文—主题概率模型以完成主题内容的识别。

2 基于引文—主题概率模型的主题识别方法

2.1 引文—主题模型

引文—主题模型是一种多重生成概率模型，如图 1 所示，其中空心点表示隐含变量，实心点表示可观测值。所使用的变量及其含义如表 1 所示。

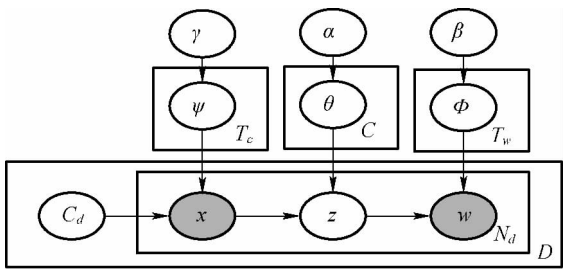

图 1 引文—主题概率模型

表 1 引文—主题模型变量及其含义

变量	含义
D，C，W，T_c，T_w	文档数、引文数、词数、引文主题数和词主题数
N_d，C_d	文档 d 中的关键词数、引文数
θ_d	主题—文档 d 生成概率
ϕ_z	词—主题生成概率
φ_z	引文—主题概率
z_n	词所属的主题
w_n，C_n	文档中的关键词、引文
α，β，γ	主题、关键词、引文的先验分布 Dirichlet 参数

假设一个文档集（corpus）R 包含 M 篇文档 $D=\{d_1, d_2, \cdots, d_M\}$ 和 N 个不同的关键词 $W=\{w_1, w_2, \cdots, w_N\}$ 以及 S 个不同的引文 $C=\{c_1, c_2, \cdots, c_S\}$。在 CTM 模型中，文档的生成是从引文集 C_d 开始。每一个主题是关键词和引文的多项式分布，一篇具有多引文和多关键词的文档在主题上的分布是引文和关键词在主题上分布的混合。对于文档 d 而言，首先从 d 的引文集 C_d 中均匀选取某个引文，然后通过对被选引文存在于的所有主题的抽样而生成关键词 w。CTM 的生成过程为：

1）$\gamma \sim \text{Dirichlet}(\beta)$，$\varphi_z \sim \text{Dirichlet}(\gamma)$，$i \in \{1, 2, \cdots, T\}$。

2）$\theta_d \sim \text{Dirichlet}(\alpha)$，$d \in \{1, 2, \cdots, D\}$。

3）文档 d 中的词 w_n 生成：选择主题 z_n，$z_n \sim \text{Multinomial}(\theta_d)$，代表当前选择的主题；选择引文 $c_n \sim \text{Unif}\{c_1, c_2, \cdots, c_S\}$；根据 $p(w_n \mid z_n, c_n; \beta, \gamma)$ 在 z_n 条件下的多项式分布，选择词 w_n。

2.2 模型参数抽样估计

CTM 模型中的参数和不能直接估计，必须采用近似推理的方法进行近似，本文采用 Gibbs 抽样方法间接求得和的值。MCMC（Markov Chain Monte Carlo）是一套从复杂的概率分布抽取样本值的近似迭代方法，Gibbs 抽样作为 MCMC 的一种简单实现形式，其目的是构造收敛于某目标概率分布的 Marlov 链，并从链中抽取被认为接近该概率分布值的样本。该方法速度快、所需内存较小、易于实现，其抽样算法过程主要有初始化、迭代和求解。

在 CTM 模型中，对于每个词 w_i、主题 z_i、引文 x_i 而言，使用基于后验概率 $p(z_i, x_i \mid w_i, C_d)$ 进行词的分布计算，其中 z_i 和 x_i 表示分配给词 w_i 的主题和引文，C_d 是文档中可观测的引文集。Gibbs 抽样估计 T 主题和 V 词汇的概率公式如下：

$$P(w_1 = m \mid x_i = k)\, p(x_i = k \mid z_i = j) \propto \frac{C_{mj}^{WT} + \beta}{\sum_{m^+} C_{m'j}^{WT} + V\beta} \frac{C_{kj}^{CT} + \alpha}{\sum_{j'} C_{kj'}^{CT} + T\alpha} \quad (1)$$

Gibbs 抽样算法如下：

1）z_i 被初始化为 1 到 T 之间的某个随机整数，i 从 1 到 N 循环，N 是文档集中的词汇个数，对 Markov 链进行初始化。

2）根据公式（1）将词汇 w_i 分配给主题 z_i，并获取 Markov 链的下一个状态。

3）迭代第 2）步，使 Markov 链接近目标分布，取 z_i（$i \in$ [1, …, N]）作为样本，并记录下来。

为了保证较小的自相关性，每迭代一定次数即记录其他样本，并舍弃当前词汇，以 w_i 表示唯一性。对于每一个单一样本，可以按照下式估算 θ、ϕ 以及 φ 的值。

$$\hat{\phi}_{mj} = \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \quad (2)$$

$$\hat{\theta}_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{d'} C_{d'j}^{WT} + T\alpha} \quad (3)$$

$$\hat{\varphi}_{cj} = \frac{C_{kj}^{CT} + \gamma}{\sum_{k'} C_{k'j}^{CT} + C\gamma} \quad (4)$$

其中，$m' \neq m$，$d' \neq d$，$k' \neq k$，α、β 和 γ 分别是主题、词和引文的 Dirichlet 分布先验参数，C_{mj}^{WT} 表示词 $w_i = m$ 分配给主题 $z_i = j$ 的次数，C_{kj}^{CT} 代表引文 $x_i = k$ 分配给主题 j 的次数。C_{dj}^{DT} 表示文档 d_i 中所有词汇被分配给主题 j 的次数。

2.3 主题识别过程

基于引文—主题模型的主题识别方法主要以文献中的关键词（包括标题关键词、摘要关键词和文献自有关键词）和引文为基础，建立文献集的引文—主题概率模型，通过多次迭代完成科技文献的主题识别过程，见图 2。

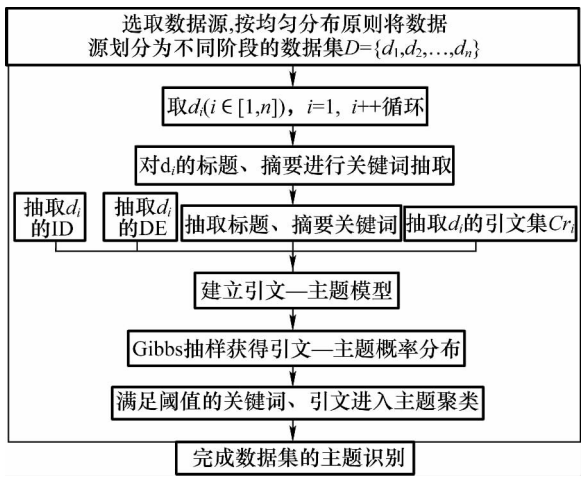

图2 基于引文—主题概率模型的主题识别过程

1) 关键词的抽取。为了完成 CTM 模型的建立，除采用文献自有关键词外，还需从文献的标题、摘要中进行关键词抽取。主要采用基于 N-Gram 和词性分析的 c_value 方法对文献的标题和摘要进行关键词抽取[9-10]，该方法在关键词抽取方面的有效性已在文献[11]中进行论述。

2) 建立文献集的引文—主题概率模型。在引文—主题概率模型中，引文和关键词共同构成文献的主题特征要素参与主题内容的识别过程。

3) 采用 Gibbs 抽样获得引文—主题概率分布。在引文—主题概率模型中，利用 Gibbs 抽样方法获得关键词、引文在潜在主题聚类上的概率分布，依次将各关键词及引文的主题概率按照降序排列，选取满足阈值的关键词和引文进入主题聚类。

3 实验及结果分析

笔者使用 Java 和 Sql Server 2005 设计一个测试系统，以完成如图 2 所示的科技文献主题识别流程。

3.1 数据集

选择“纳米轻质材料”技术领域中主题为“碳纳米纤维”的科技论文作为实证研究对象。考虑到数据的完整性和权威性，以及引文数据的可获得性，选择 ISI Web of Knowledge 为数据检索平台，检索路径为：{ TS = (("* carbon fiber* " or "* carbon fibre* ") and ("* nano* ")，数据库 = “SCI-EXPANDED，SSCI，CPCI-S，BK-CI-S，BKCI-SSH，CCR - EXPANDED，IC" ，时间跨度 = 所有年份}。检索时间：2012 年 9 月 10 日，共返回 2 727 条记录。本文选取 2002—2011 年间的 10 年论文作为主题识别的数据源，共 2 120 篇论文，并分别抽取论文关键词、标题关键词、摘要关键词和引文。在经过关键词、标题、摘要、引文的论文特征项提取，并对论文的标题、摘要进行关键词抽取、停用词过滤、词干还原、同义词转化、术语识别等必要的数据处理和清洗，得到的结果见表 2。

表2 论文数据统计表（2002—2011 年）

时间	论文量	自有关键词数量	标题关键词数量	摘要关键词数量	参考论文数量
2002	90	766	313	3672	2314
2003	107	802	315	4169	2658
2004	120	1092	389	5435	3249
2005	180	1638	540	8589	5269
2006	160	1367	492	7413	3825
2007	234	2042	849	11023	5853
2008	243	2297	887	11609	6820
2009	282	2728	1112	13550	8252
2010	323	3159	1243	15321	10649
2011	381	3735	1586	18749	11924
合计	2120	19626	7726	99530	60813

3.2 模型实现

本文使用 C + + 语言实现 CTM 概率模型，参照文献[2]给出的模型参数值，实验模型的参数值赋值为：$\alpha = 0.5$，$\beta = 0.1$，$\gamma = 0.1$，迭代次数 = 1000，主题数 = 20。经过多次迭代后生成如表 3 中所示的主题信息内容，其中每个主题中包含 30 个关键词及其概率分布，以及 30 个引文及其概率分布，分别保存在参数 ϕ 和 φ 中。对 ϕ 和 φ 做进一步细致的分析将有助于论文主题特征的准确识别。

表3 2#主题的主题分布部分数据（2011 年）

关键词	词—主题概率
carbon nanotube	0.02929
carbon fiber	0.02104
carbon nanotube cnt	0.00947
mechanical property	0.00610
Composite	0.00493
tensile strength	0.00483
fiber using	0.00341
interfacial shear strength	0.00326
surface carbon fiber	0.00295
interfacial shear	0.00288
……	……
引文	**引文—主题概率**
sager rj，2009，v69，p898，compos sci technol	0.00833
thostenson et，2002，v91，p6034，j appl phys	0.00777
qian h，2010，v20，p4751，j mater chem	0.00722
iijima s，1991，v354，p56，nature	0.00679
chou tw，2010，v70，p1，compos sci technol	0.00580
qian d，2000，v76，p2868，appl phys lett	0.00529
zhu s，2003，v12，p1825，diam relat mater	0.00470
de riccardis mf，2006，v44，p671，carbon	0.00427
zhang fh，2009，v44，p3574，j mater sci	0.00410
he xd，2007，v45，p2559，carbon	0.00407
……	……

3.3 结果分析

对“纳米碳纤维”技术中 2002—2011 年间的论文集建立引文—主题概率模型，其中 θ 中保存着所有的主题—文档概率分布，ϕ 中保存着所有的关键词—主题概率分布，φ 中保存着所有引文—主题概率分布，对这 3 个参数

中所包含的概率分布做进一步分析，将有助于主题特征的进一步明确，并为下一步主题的演化分析打下坚实的基础。

以该技术领域中的“carbon fiber”（碳纤维）主题为例，在2011年论文中的20个主题中，“carbon fiber”关键词出现了20次，也即本年度所有20个主题中都出现了该关键词，且该关键词以不同的概率分布于不同的主题，具体如表4所示。

表4 “carbon fiber”主题概率分布（2011年）

关键词	主题编号	词—主题概率	主题编号	词—主题概率
carbon fiber	2	0.02104	17	0.00689
	1	0.01629	13	0.00512
	10	0.01627	4	0.00265
	12	0.01583	18	0.00195
	15	0.01559	8	0.00179
	11	0.01400	9	0.00176
	3	0.01329	6	0.00154
	20	0.01232	14	0.00153
	7	0.01055	5	0.00091
	16	0.00781	19	0.00037
	……	……	……	……

从表4可以看出，“carbon fiber”关键词在2#主题中的分布概率最高，因此可以选择2#主题作为“carbon fiber”技术的主要分析对象。再对θ进行统计分析，可以得到2#主题的主题—文档概率分布，具体数据如表5所示。

表5 2#主题的文档概率分布（2011年部分数据）

主题编号	论文编号	主题—文档概率
2	23	0.74343
	52	0.65800
	186	0.64583
	259	0.63309
	165	0.61183
	292	0.60466
	181	0.58292
	372	0.50211
	83	0.49814
	45	0.49225
	……	……

从表5所显示的2#主题在2011年的381篇论文的概率分布排序结果中可以看出，该主题在23#论文中的分布概率最高，因此可以判定，在2011年，“carbon fiber”技术研究的主题分析对象是2#主题，论文分析对象是23#论文。2#主题作为“carbon fiber”技术的重点分析对象，该主题中所包含的关键词及其概率分布、引文及其概率分布，具体如表3所示。表3中诸如“sager rj，2009，v69，p898，compos sci technol”、“thostenson et，2002，v91，p6034，j appl phys”、“qian h，2010，v20，p4751，j mater chem”等高概率分布的引文在其他主题中也曾出现，这样，带有引文信息的主题将为下一步主题演化分析提供一定的分析基础。

基于引文—主题概率模型的主题识别方法能够获得科技文献中更完整的主题内容，这将为主题进一步演化分析提供基础。同时，在2010年和2011年引文中，共有25篇引文在这两年中都被引用，因此，将引文引入主题可以为主题进一步演化分析提供重要的量化分析作用。

4 结束语

科学文献中的主题内容通常都是由文献中包含的关键词所描述。为一篇文献建立主题分布模型时常用的方法就是把每个主题看成是词的概率分布，而把文献视为这些主题的概率分布。本文在贝叶斯概率模型中引入引文因素，将引文作为模型的一项参数，与关键词共同建立引文—主题模型。实验表明，该模型能较全面、深入地识别科技文献的主题内容。但是，由于引文—主题模型计算量较大，所以文献集的数据清洗直接影响到模型的效率，因此下一步的研究重点应该是如何优化数据清洗方法，提高标题关键词和摘要关键词抽取的准确性和效率。同时，该模型的有效性有待于在更多的技术领域中进行验证。□

参考文献

[1] BLEI D M，et al. Latent dirichlet allocation [J]. Journal of Machine Learning Research，2003（3）：993-1022.

[2] GRIFFITHS T L，STEYVERS M. Finding scientific topics [C] //Proceedings of the National Academy of Sciences of the United States of America，2004：5228-5235.

[3] 石晶，李万龙. 基于LDA模型的主题词抽取方法 [J]. 计算机工程，2010，36（19）：81-83.

[4] 王金龙，徐从富，耿雪玉. 基于概率图模型的科研文献主题演化研究 [J]. 情报学报，2009，28（3）：347-355.

[5] BOLLEI L，et al. Finding topic trends in digital libraries [C] //Proceedings of the 2009 ACM/IEEE Joint Conference on Digital Libraries. Austin，Texas，USA，2009：69-72.

[6] BOLELLI L，ERTEKIN S，GILES C L. Topic and trend detection in text collections using latent dirichlet allocation [C] //Proceedings of the 31st European Conference on Information Research. Toulouse，France，2009：776-780.

[7] 王萍. 基于概率主题模型的文献知识挖掘 [J]. 情报学报，2011，30（6）：583-590.

[8] YOOKYUNG J，CARL L，et al. Detecting research topics via the correlation between graphs and texts [C] //Proceedings of 13th International Conference on Knowledge Discovery and Data Mining，2007：370-379.

[9] 王小捷，常宝宝. 自然语言处理技术基础 [M]. 北京：北京邮电大学出版社，2002.

[10] FRANTZI K，ANANIADO S，TSUJII J. The c-value/NC-value method of automatic recognition for multi-word terms [C] //Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries，1998：585-604.

[11] 叶春蕾，冷伏海. 科技文献全文主题识别方法实证研究 [J]. 现代图书情报技术，2012（1）：53-57.

作者简介：叶春蕾，女，1975年生，博士，副教授。
冷伏海，男，研究员，博士生导师。

收稿日期：2013-03-18