© 山东大学科技期刊社版权所有

文章编号:1671-9352(2018)09-0023-12 DOI:10.6040/j.issn.1671-9352.1.2017.044

基于复合主题演化模型的作者研究兴趣动态发现

余传明[1],左宇恒[1],郭亚静[1],安璐[2*]
(1. 中南财经政法大学信息与安全工程学院,湖北 武汉 430073;
2. 武汉大学信息管理学院, 湖北 武汉 430072)

摘要:以金融领域的科技文献作为实验数据,提出了一种新的用于动态挖掘领域相关的作者研究兴趣的复合主题演化模型。该模型能够获取作者在不同时间片下的主题概率分布以及主题下词汇概率分布,并充分考虑作者在合作作者文献中的排名对于其研究主题和主题变化的影响。通过金融领域的实证研究表明,该复合主题演化模型能够有效地揭示金融领域作者研究兴趣的动态变化。

关键词:主题演化模型;复合主题演化模型;主题挖掘
中图分类号:TP391 **文献标志码**:A
引用格式:余传明,左宇恒,郭亚静,等. 基于复合主题演化模型的作者研究兴趣动态发现[J]. 山东大学学报(理学版),2018,53(9):23-34.

Dynamic discovery of authors' research interest based on the combined topic evolutional model

YU Chuan-ming[1], ZUO Yu-heng[1], GUO Ya-jing[1], AN Lu[2*]
(1. School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, Hubei, China;
2. School of Information Management, Wuhan University, Wuhan 430072, Hubei, China)

Abstract: We propose a new combined topic model, i.e. author topic time-latent dirichlet allocation(ATT-LDA) with author ranking (AR), for the of dynamic discovery of researchers' interest, which is based on the academic literature in the financial field. Through the proposed model, we can easily acquire the probability distribution of the authors' interest, as well as the probability distribution of topics on deferent words. The influence of the ranking in the co-author list are fully taken into consideration. The empirical study shows that the proposed method can effectively reveal the dynamic change of interest of the authors in the financial field.
Key words: topic evolution model; combined topical model; topic mining

0 引言

科技文献是科学技术人员经过科学实验研究,对自然、社会科学或者工程技术等领域内的问题进行分析研究,反映领域内的新理论、新思想、新技术的学术论文。科技文献的海量增长给研究学者在进行学术交流和学习中高效获取有用信息提出了棘手的挑战。从科研人员的学术研究行为来看,在其研究某领域的主题之前,通常需要了解该主题在某时间段内的主要内容,以及该主题的研究内容随时间的变化等。对某一领域内各个主题的研究内容以及各个主题的研究发展趋势进行手工获取和分析需要花费科研工作者较多的精力,如果借助机器学习、人工智能等辅助手段,则能有效节省科研工作者的信息获取时间。

收稿日期:2017-07-04; **网络出版时间**:2018-04-19 16:10:57
网络出版地址:http://kns.cnki.net/kcms/detail/37.1389.N.20180419.1330.002.html
基金项目:国家自然科学基金资助项目(71373286);教育部哲学社会科学研究重大课题攻关项目(17JZD034);国家自然科学基金青年科学基金项目(71603189)
第一作者简介:余传明(1978—),男,博士,副教授,研究方向为数据挖掘与商务智能. E-mail:yucm@ zuel.edu.cn
*通信作者简介**:安璐(1979—),女,博士,教授,博士生导师,研究方向为可视化知识发现. E-mail:anlu97@ 163.com

针对上述问题,本文以金融领域的科技文献作为实验数据,提出了一种新的用于动态挖掘领域相关的作者研究兴趣的 ATT-LDA(author topic time-latent dirichlet allocation with author ranking)复合主题演化模型。该模型能够获取作者在不同时间片下的主题概率分布以及主题下词汇概率分布,并充分考虑作者在合作作者文献中的排名对其研究主题和主题变化的影响,能够较好地挖掘作者在连续的时间段内的研究兴趣以及随着时间变化其兴趣强度的变化趋势,从而让科研人员更深入、更全面地了解作者主题的发展变化,进而快速地找到适合的科研合作者或为其科研决策提供帮助等。

1 相关工作

在主题建模领域,最为经典的方法是 Blei 等[1]提出的 LDA (latent dirichlet allocation)模型,该模型主要用于对文档进行主题建模。Rosen-Zvi 等[2-3]提出了 AT(author-topic)模型,将科研人员所对应的作者-主题分布应用于 LDA 模型中,从而发现作者的主题信息也就是作者的研究兴趣。该模型的缺陷在于其首先假定每个作者只研究一个主题。为了克服这一限制,Mimno 等[4]在 AT 模型的基础上构建了 APT (author-persona-topic)模型。该模型将科技文献分为多类,这些类代表该科研人员不同的"身份(persona)",并给出了一种估计研究兴趣个数的启发式方法。在此基础上,Kawamae[5-6]对 APT 的数据来源进行扩充,提出了一种以文档类为基础的 AIT(author-interest-topic)模型。在国内,李纲等[7]从作者微观个体研究兴趣角度出发,通过对作者合著网络中作者关联关键词集的研究,定量地验证研究兴趣相似是作者合作的一个动机。徐戈等[8]对于主题模型在自然语言处理中的发展进行了详细的介绍和分析,有助于我们更深刻理解主题模型发展中各项工作的联系。为了更准确、更合理地进行科研合作推荐,刘萍等[9]运用社会网络理论,在社区划分的基础上,构建基于 LDA 的作者兴趣模型,然后对作者的相关文献进行分析,实现科研合作推荐的目的。Han[10]提出了一个改进的 LDA 模型,应用于新闻媒体领域中,从社会媒体评论中提取出隐含的主题类别体,用以描述用户的兴趣和兴趣热点内容,该方法能够有效提高个性化的社会化媒体服务。余传明等[11]提出一种结合 LDA 主题模型和自然语言处理技术的方法用以挖掘用户评论热点。陈文涛等[12]分析了不同的主题模型,通过实验比较了 3 种主题模型构建微博用户兴趣模型的性能。

在作者兴趣的动态演化,较多学者从时间变化趋势、地理变化趋势等多角度进行研究。通常而言,对于作者兴趣随时间变化的分析方法有 3 种[13]:第 1 种是将时间因素作为主题模型的内在变量;第 2 种是先获取主题再离散到时间窗;第 3 种是先离散到时间窗再获取主题。Blei 等[14]借助时间序列分析方法构建了 DTM(dynamic topic model)模型,该方法在时间切片的基础上进行主题建模。由于先对数据集进行了切分,其后进行的主题发现实验所获得的主题内容并不是一致的,因此相邻时间片间的主题难以进行追踪。另外,该方法中的时间粒度的选择也会对实验结果产生较大影响。在此之后,Wang 等[15]和 Xu 等[16]在 DTM 模型的基础上分别提出了 cDTM 模型以及 ToT(topic over time)模型。倪丽萍等[17]针对主题演化分析中阈值设定和主题漂移的问题,提出一种 LDA-AP 主题演化模型,利用 AP 聚类算法对不同时间窗口内的多个主题进行聚类,其中计算主题相似度采用加入时间衰减因子的 JS 散度来度量,最后对多个主题内容进行演化分析。与此相关的研究的还包括作者-主题模型以及相关主题模型等。作者-主题模型 (author-topic model, ATM)[2-3]利用话题对文章作者和文章内容建模,话题分布受到作者分布的影响,CTM(correlated topic model)[14]不仅考虑文档之间的关系,还引入了话题间的关系。史庆伟等[18]针对挖掘大规模科技文献中作者、主题和时间及其关系的问题,考虑科技文献的内外部特征,提出了一个作者主题演化(AToT)模型。廖君华等[19]开发了一套分析主题变化的系统,实现了网络热点话题数据源自动定制、自动获取和自动存储。针对中文话题分析,李保利等[20]提出了一种考虑无意义话题对于主题演化的负面影响的主题模型。这种方法首先利用 LDA 模型对已经进行时间切片的数据集进行话题抽取,得到不同数量的话题,并借助简单启发式规则从主题强度和内容进行分析,以得到研究主题的演化趋势。

基于上述研究方法,本文在 TOT 模型基础上加以改进,提出了作者-主题分布的 ATT 模型,同时在 LDA 模型的基础上加入作者在合作者文献中的排名因素得到 LDA(AR)模型,并将这 2 种模型相结合,通过 ATT-LDA(AR)复合主题模型来研究金融领域内的作者研究兴趣在不同时间切片下的变化情况。

2　作者研究兴趣建模

2.1　ATT 模型

将时间因素引入模型,在 TOT 模型[16]的基础上将“作者-词汇”矩阵取代“文档-词汇”矩阵形成一个“作者-主题”演化模型 ATT(auhtor-topic-time)模型,ATT 模型概率图如图 1 所示,符号说明见表 1。在 ATT 模型中,将同一个时间段下的文档合并为一篇文档,每位作者在一个时间片下只对应一篇文档,将作者-词汇矩阵作为输入应用于 ATT 模型。从图 1 中可以看出,主题 z 下的词汇分布由文档下的词汇以及时间共同决定。

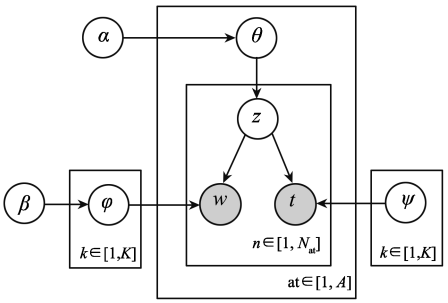

图 1　ATT 模型
Fig.1　ATT model

表 1　符号说明
Table 1　Symbol description

符号	描述
K	主题的数量
A	作者的数量
V	词项的数量
N_{at}	作者 at 文档中单词的数量
ϑ_{at}	作者 at 上的主题多项分布,$\Theta=\{\vartheta_{\mathrm{at}}\}_{\mathrm{at}=1}^{A}$($A\times K$ matrix)
φ_k	主题 k 上的词项多项分布,$\Phi=\{\varphi_k\}_{k=1}^{K}$($K\times V$ matrix)
ψ_k	主题 k 随时间变化的贝塔分布,$\Psi=\{\psi_k\}_{k=1}^{K}$($K\times 2$ matrix)
$z_{\mathrm{at},N}$	作者 at 文档中第 n 个词的主题
$w_{\mathrm{at},N}$	作者 at 文档中第 n 个词
$t_{\mathrm{at},N}$	作者 at 文档中第 n 个词的时间戳
$\boldsymbol{\alpha}$	作者多项分布 ϑ_{at} 的狄利克雷先验分布,K 维向量
$\boldsymbol{\beta}$	主题多项分布 φ_k 的狄利克雷先验分布,V 维向量

在图 1 中,某一词汇的产生过程的步骤如下:

1) 对于每个主题 $k\in[1,K]$,依据狄利克雷先验 β 抽取多项分布 $\varphi_k\sim\mathrm{Dirichlet}(\beta)$;

2) 对于每个作者 $\mathrm{at}\in[1,A]$,依据狄利克雷先验 α 抽取多项分布 $\vartheta_{\mathrm{at}}\sim\mathrm{Dirichlet}(\alpha)$;

3) 对于每个作者的文档中的每个单词 $n\in[1,N_{\mathrm{at}}]$,(a)抽取一个主题,$z_{\mathrm{at},n}\sim\mathrm{Multinomial}(\vartheta_{\mathrm{at}})$,(b)抽取一个单词,$w_{\mathrm{at},n}\sim\mathrm{Multinomial}(\varphi_{z_{\mathrm{at},n}})$;(c)抽取一个时间戳 $t_{\mathrm{at},n}\sim\mathrm{Beta}(\psi_{z_{\mathrm{at},n}})$。

如上述过程所示,主题的后验分布取决于作者文档和时间两种信息。ATT 模型需要计算的参数:$\vartheta_{\mathrm{at}}|\boldsymbol{\alpha}\sim\mathrm{Dirichlet}(\boldsymbol{\alpha})$,$\varphi_k|\boldsymbol{\beta}\sim\mathrm{Dirichlet}(\boldsymbol{\beta})$,$z_{\mathrm{at},n}|\vartheta_{\mathrm{at}}\sim\mathrm{Multinomial}(\vartheta_{\mathrm{at}})$,$w_{\mathrm{at},n}|\varphi_{z_{\mathrm{at},n}}\sim\mathrm{Multinomial}(\varphi_{z_{\mathrm{at},n}})$,$t_{\mathrm{at},n}|\psi_{z_{\mathrm{at},n}}\sim\mathrm{Multinomial}(\psi_{z_{\mathrm{at},n}})$。通过直接计算的方法是无法得到这些参数值的,可通过抽样的方法进行简单地估计。

2.2　LDA(AR)模型

在文献数据中,25.6%的文献为多个作者合作发表,而作者排名也会体现出其研究兴趣。基于此,本文提出 LDA(AR)(LDA-author ranking)模型,该模型首先使用 LDA 主题模型对整个文献集挖掘主题信息,后离散到时间窗并加入了作者的排名信息的方式,获取作者下的主题分布,分析在排名信息影响下作者主题强度随时间变化趋势,具体的模型流程如图 2 所示。

首先,将数据集中的文档-词汇矩阵输入 LDA 主题模型中,通过计算可以获取到文档-主题概率分布(ϑ_d),主题-词项概率分布(ϕ_T),然后根据设置好的时间切片方式和文档标记的时间戳,使文档落到不同的时间窗中。为了获取各个时间段内作者-主题概率分布,整合作者在同一个时间段内的文档-主题概率分布(θ_A),并在文档-主题概率分布中加入作者排名特征,即公式:

$$\theta_A = \frac{\sum_{1}^{m} \lambda_{m,t}\theta_{d_{m,t}}}{m}, \tag{1}$$

其中,m 为该作者所发表的文献数量,t 为某一个时间段,λ 为作者排名的权重信息,其计算公式为

$$\lambda_{m,t,r} = \frac{n + 1 - r}{n}。 \tag{2}$$

其中,n 为该作者的第 m 篇文献中合作作者个数,r 为该作者在此篇文献中的排名。

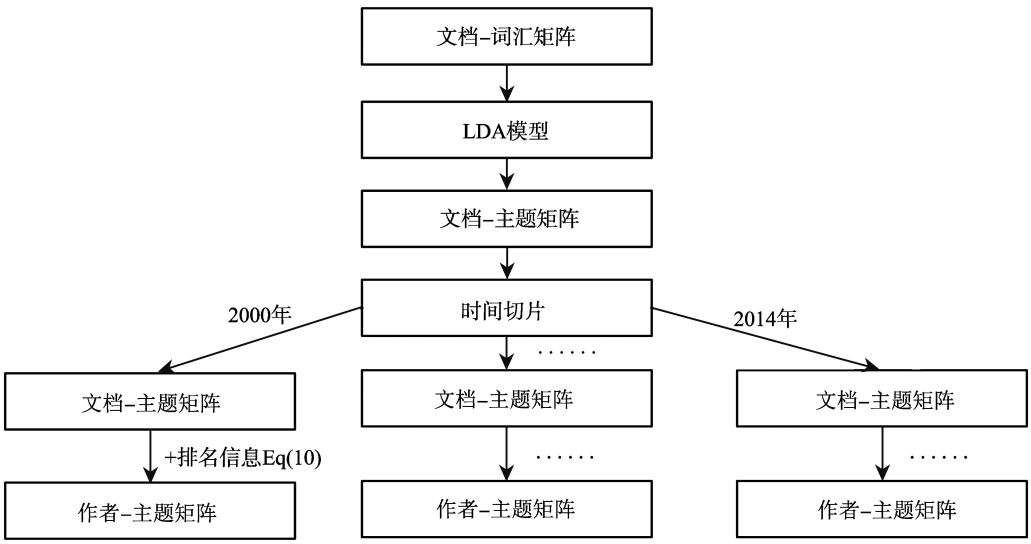

图 2 LDA(AR) 模型流程图

Fig.2 Flow chart of LDA(AR) model

公式(1) 和(2) 的思想是排名越靠后的作者对该文档的主题兴趣越弱,而对排名第一的作者,本文认为该文献能够很好地表达该作者的主题兴趣,故权值为 1,排名越后的作者权值越小,排序为 $i + 1$ 的作者兴趣度权重比排序为 i 的作者兴趣度权重低 $1/n$。本文所提出的算法的目的是测量作者在各主题上的兴趣度,不同于作者对文献的贡献度,模型中同一论文的各个作者的权重之和不一定等于 1。

由于 ATT 模型的输入是时间 - 作者 - 词汇矩阵,一个作者在同一个时间下的文档被合并成一个词汇集,故而失去了作者排名的因素,无法实现在 ATT 模型中加入作者排名权重,因此,在 LDA 模型中加入作者权重后和 ATT 模型组合成一个复合模型。

2.3 复合主题演化模型

为了获取作者在连续的时间段内研究主题的概率分布,同时考虑到时间因素和作者排名对主题分布的影响,本文构建了复合主题演化 ATT + LDA(AR) 模型,通过该模型计算作者 - 主题概率分布。其原理就是对原始数据进行预处理获取词向量,然后分别利用 ATT 模型和 LDA(AR) 模型获得作者 - 主题 - 时间概率分布矩阵、主题 - 词项概率分布矩阵;将上一步的输出通过主题对齐和整合,得到一组作者 - 主题 - 时间概率分布矩阵、主题 - 词项概率分布矩阵,作为最后结果输出作为最终实验结果。

研究表明,尽管通过动态主题建模得到的主题在时间轴上具有局部性,即某些主题局限于某个短的时间内,但大多数主题在时间序列下相对稳定,极少主题出现完全转换成为另一个主题的情况[21]。为了探究这些相对稳定的主题在数十年间的演化情况,我们尝试将 ATT 模型与 LDA-AR 模型的主题进行对齐,然后对主题进行动态演化分析。具体实现流程如图 3 所示。

特征选择在本文中为特征词的选择,即研究作者兴趣特定的一些词,这些词区别于别的研究,特指某个研究领域,比如金融方向的“投资”“市场”“银行”“风险” 等等。特征权值计算在本文中为词频向量的生成,即权值由该词出现的次数(词频) 决定。

由于 2 个模型所计算的主题并不是完全一致的,因此必须通过主题对齐的步骤才能将 2 个模型所产生的主题一一映射,在这里采用计算相似度的方式将 2 个相似度最高的主题对应起来。采用计算相似度的方

法对主题进行追踪对应最早是由 Blei 等[14] 中所提出的 DTM(dynamic topic model) 模型中提出的,DTM 模型在第一章节中有详细描述。DTM 模型首先将文档集中的文档根据其自带的时间标记切分为多个子数据集,通过这些预处理的工作之后,设置一个确定的主题数量 K,再分别对每个子数据集使用 LDA 模型进行处理获取主题信息。通过这样的方式可以获取到不同时间片下的主题信息,然而这样会导致每个时间片下主题的内容信息也就是主题下的词汇内容并不一致,这就是主题不对齐的情况,Blei 等[14] 使用 KL 距离(kullback-leibler divergence) 计算不同时间窗内主题分布的相似程度。

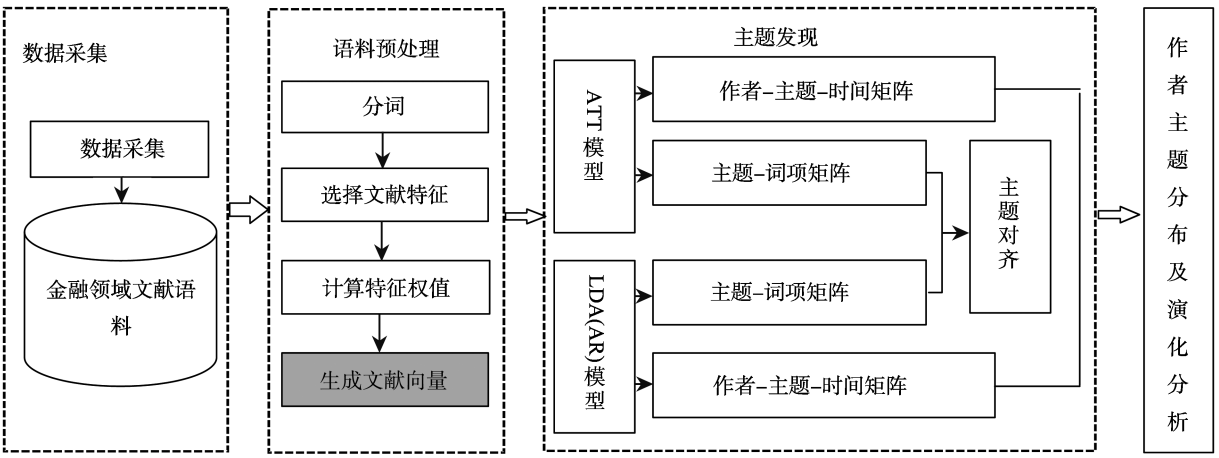

图 3　ATT + LDA(AR) 模型流程
Fig.3　Flow chart of ATT + LDA(AR) model

通常,对于在主题对齐步骤中所使用的相似度算法有 2 种:一是文本相似度计算;二是 Blei 所使用的这种概率分布相似度计算方法。在本文的主题对应过程中,需要通过判断主题下词汇的概率分布的相似程度来进行主题对齐,根据实验所需,采用 KL 距离算法来计算不同主题 - 词项分布之间的相似度。KL 距离算法如公式(3) 所示:

$$\boldsymbol{D}(P \parallel Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}, \tag{3}$$

其中,$\boldsymbol{D}(P\|Q)$ 表示为 Q 到 P 的 KL 距离,X 为事件,$P(x)$、$Q(x)$ 分别为 P 和 Q 的概率分布。推广到本文的主题相似度计算中如公式(4) 所示:

$$\boldsymbol{D}(\mathrm{Topic}_1 \| \mathrm{Topic}_2) = \sum_{v \in V} \mathrm{Topic}_1(v) \log \frac{\mathrm{Topic}_1(v)}{\mathrm{Topic}_2(v)}, \tag{4}$$

其中,$\mathrm{Topic}_1 = \{z_1, z_2, \cdots, z_k\}$ 表示为 ATT 模型所产生的主题;$\mathrm{Topic}_2 = \{z_1, z_2, \cdots, z_k\}$ 表示为 LDA(AR) 模型所产生的主题;$v \in V$ 表示词汇;$\boldsymbol{D}(\mathrm{Topic}_1 \| \mathrm{Topic}_2)$ 为 ATT 模型与 LDA(AR) 模型的主题之间的 KL 距离矩阵,该矩阵大小为 $K \times K$。

在计算相似度的方法将主题一一映射后,通过公式将其整合起来:

$$\vartheta_A = \rho \vartheta_{1A} + (1 - \rho)\, \vartheta_{2A}, \tag{5}$$

其中,ϑ_{1A} 为 ATT 模型得到的作者 - 主题 - 时间概率分布矩阵;ϑ_{2A} 为 LDA(AR) 模型得到的作者 - 主题 - 时间概率分布矩阵;ρ 为一个[0,1] 之间的一个值,用于调节 2 个模型的比例。本实验取 $\rho = 0.5$,即综合考虑 2 个模型中的排名特征以及时间因素,二者对复合模型的贡献度一样大,均为 1/2。

3　实证结果与分析

3.1　数据集

本文实验采用的数据集是从网络上抓取的 2000—2014 年金融领域的 68 905 篇期刊论文, 其中作者共有 45 567 名。表 2 展示了 2000—2014 年内所抓取的数据中每年期刊文献数量,表中显示从 2000—2007 年每年期刊文献数量平均为 3 000 篇左右,在 2008—2014 年每年期刊文献数量平均为 5 800 篇左右。

表 2 金融领域期刊文献年度分布表
Table 2 Annual distribution of financial periodicals

年份	文档篇数	总占比/%	年份	文档篇数	总占比/%
2000	2 736	4.09	2 008	5 793	8.66
2001	3 176	4.75	2 009	5 862	8.77
2002	3 051	4.56	2 010	5 862	8.77
2003	3 150	4.71	2 011	5 818	8.70
2004	3 184	4.76	2 012	5 801	8.68
2005	3 324	4.97	2 013	5 816	8.70
2006	3 369	5.04	2 014	5 794	8.67
2007	4 123	6.17			

3.2 主题困惑度计算

困惑度(perplexity) 是 LDA 模型等主题模型的一个评价标准,其定义可以理解为对于一篇文档 d,主题模型对该文档属于哪个主题的不确定程度就是困惑度。在其他条件固定的情况下,困惑度越小,主题的确定程度越高,但是容易过拟合。Blei 等[1] 中给出如下方法计算困惑度:

$$\text{perplexity}(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right\}, \tag{6}$$

其中,N_d 为文档 d 的词汇数量(不去重);w_d 为文档 d 中的词汇;对于文档中的每一个词汇 w,其概率 $p(w) = \sum_z p(z \mid d) \times p(w \mid z)$,也就是该词在所有主题中概率分布的值乘以该词所属的主题在文档中的概率。

类比LDA 模型,ATT模型在计算主题困惑度时,用作者 - 主题概率矩阵替换LDA模型主题困惑度计算公式中的文档 - 主题概率矩阵。其中 N_d 为作者 a 的词汇数量(不去重);w_d 为作者 a 中的词汇;对于作者中的每一个词汇 w,其概率 $p(w) = \sum_z p(z \mid a) \times p(w \mid z)$,也就是该词在所有主题中概率分布的值乘以该词所属的主题在作者中的概率。

为了确定本文所涉及的 ATT 模型与 LDA(AR) 模型实验中的主题数 k,将参数 α 设置为 $50/k$,β 设置为 0.01,迭代次数为 1 000 的条件下,选定不同的 k 值,分别计算 ATT 模型和 LDA(AR) 模型的困惑度值进行分析确定实验中的主题数。

图 4 展示的是保持其他参数固定不变的情况下,主题模型的主题数 k 选择从 10 开始逐渐增大到 100 时,ATT 模型与 LDA 模型的困惑度的值,整体展示的是 ATT 模型与 LDA(AR) 模型的困惑度变化趋势。从图 4 中可以看出,当 k =30 时,LDA(AR) 模型的困惑度最低,为 194.993,同时 ATT 模型的困惑度为227.383,此时 LDA 模型与 ATT 模型的困惑度都为最低值。根据以上分析,本文实验中 LDA(AR) 模型与 ATT 模型的主题数选定为 30。

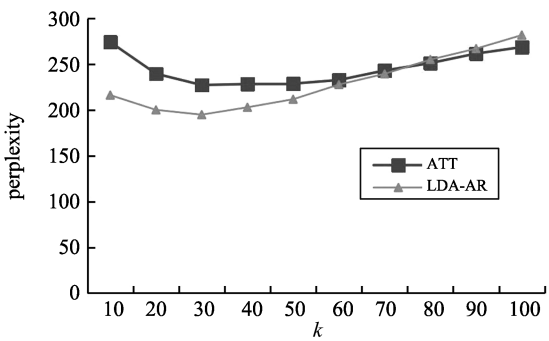

图 4 ATT 和 LDA(AR) 模型主题数 - 困惑度
Fig.4 ATT and LDA(AR) model topic number-perplexity

3.3 主题对齐结果分析

确定 k = 30,其他参数与上节困惑度实验中一致,分别计算出 ATT 模型与 LDA(AR) 模型中的作者 - 主题概率分布矩阵和主题 - 词汇概率分布矩阵,但是这 2 个模型所计算的主题并不是一一对应的,需要通过主题对齐的步骤使之映射起来。

在本文的数据集上,这 2 种映射得到的主题匹配并不相同。由表 3 可以看出,Topic 1(ATT) 到 Topic 2(LDA-AR) 的距离为 5.5,Topic 2(ATT) 到 Topic 1(LDA-AR) 的距离为 6.6,二者并不相同。对比行与列的结果可以看出,在 LDA-AR 的所有主题中,与 ATT 的主题 i 最近的是主题 j,那么在 ATT 的所有主题中,与

LDA-AR 的主题 j 最近的不一定是主题 i。由于 ATT 模型中主题的概率分布受限于时间,其主题词汇概率分布具有时间局限性,为了探究相对稳定的主题在数十年间的演化情况,本文以 LDA-AR 模型的主题词汇概率分布为主,使用 ATT 到 LDA-AR 的主题映射,即选取列中的最小值。

表 3　ATT 模型到 LDA-AR 模型的 KL 距离计算结果(部分)

Table 3　The calculation result of KL distance from ATT to LDA-AR model(part)

ATT	LDA-AR														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	6.1	5.5	10.9	6.0	6.1	6.4	5.9	6.2	6.4	6.3	6.1	5.8	13.3	6.2	13.3
2	6.6	6.5	12.9	6.5	6.6	6.4	6.7	6.4	6.8	6.8	6.7	**5.2**	13.7	6.8	13.1
3	5.8	7.1	9.8	7.2	7.2	6.2	7.0	7.2	7.2	7.2	7.1	7.1	13.3	7.2	**9.8**
4	5.4	4.9	11.9	6.6	6.1	6.5	6.6	6.5	6.3	6.6	6.4	6.0	11.0	6.5	13.5
5	6.1	6.1	12.7	5.9	**5.8**	6.2	6.1	5.9	6.1	4.8	5.9	5.5	12.7	5.0	12.8
6	6.6	7.1	13.5	7.3	6.8	5.4	7.3	7.3	7.3	7.3	7.1	6.9	**8.8**	7.2	10.1
7	6.9	6.9	13.6	6.7	6.8	6.9	6.8	6.1	6.6	6.8	6.9	5.3	13.6	6.7	13.1
8	7.0	6.8	14.1	6.9	6.2	7.1	7.1	5.9	6.6	7.1	7.0	6.5	14.2	**3.4**	13.9
9	7.0	6.6	12.9	6.7	6.9	6.5	6.9	**4.2**	7.0	5.8	7.0	6.9	11.9	6.8	13.7
10	6.5	6.5	13.2	6.1	6.4	6.5	6.5	6.4	6.2	5.4	5.9	6.1	13.4	6.1	11.8
11	8.2	7.3	13.5	8.2	8.2	8.2	6.1	7.5	4.3	8.1	8.2	8.1	14.2	8.1	13.8
12	7.1	7.1	13.7	7.1	7.1	7.2	6.9	5.4	6.6	7.2	7.0	7.1	14.0	7.2	13.7
13	7.8	5.8	10.6	6.7	7.3	4.3	7.2	7.8	7.8	7.8	7.8	7.6	10.7	7.7	12.9
14	7.2	7.2	14.0	**3.6**	6.5	7.2	7.0	7.2	7.1	6.6	7.0	6.6	14.3	7.1	13.1
15	6.1	5.6	12.5	6.0	6.0	6.0	**5.8**	6.2	6.1	5.5	**5.5**	6.0	13.0	6.2	12.6
16	7.1	4.4	11.3	6.8	7.2	6.6	6.8	7.3	7.3	7.2	6.4	6.5	14.0	7.1	13.9
17	7.3	6.0	13.4	6.6	6.6	7.2	7.1	6.6	5.6	6.5	6.8	7.2	13.2	7.1	13.3
18	5.0	6.2	10.3	6.7	7.9	5.3	7.6	8.0	7.8	7.7	7.8	7.8	10.2	8.0	14.1
19	8.0	5.8	9.3	6.4	8.0	5.2	8.0	7.9	7.8	8.0	7.8	7.9	14.1	8.0	14.2
20	4.4	6.5	10.1	6.9	7.6	5.2	7.7	7.8	7.5	7.8	7.8	7.8	9.5	7.2	12.0
21	7.9	7.3	12.7	7.9	6.9	7.9	7.3	6.5	**3.1**	7.5	7.6	7.9	14.9	7.9	14.2
22	6.9	6.9	13.1	6.9	6.8	7.0	6.9	6.9	7.0	6.9	5.9	7.0	13.7	7.0	13.6
23	7.3	5.1	**9.1**	5.6	7.2	**4.0**	7.3	7.0	7.1	7.0	6.4	7.2	13.3	7.0	13.7
24	6.9	**4.3**	11.4	7.0	7.4	7.5	6.2	7.3	7.3	7.5	7.4	7.3	13.7	6.9	11.2
25	6.8	6.4	13.8	6.4	6.6	6.8	6.1	6.0	6.4	5.2	6.7	6.7	13.8	6.7	13.5
26	**4.0**	5.6	13.0	6.3	6.6	6.2	6.4	6.4	6.7	6.6	6.7	6.4	11.2	6.7	13.5
27	6.4	6.8	12.2	6.3	6.3	6.8	6.6	5.8	6.6	6.7	6.4	6.2	13.4	6.2	13.2
28	7.1	6.6	12.9	7.1	7.1	7.1	6.9	7.0	6.6	7.1	7.1	5.9	12.1	7.1	13.1
29	6.9	6.8	13.4	6.7	6.6	6.9	6.8	6.7	6.8	**3.3**	6.5	6.6	13.7	6.8	13.8
30	6.2	6.0	12.3	6.0	6.1	6.2	6.2	5.4	5.3	6.0	6.2	5.7	13.0	6.2	13.0

将 2 种模型的主题一一对应时,我们考虑以下 3 种规则:

(1)选取 ATT 模型主题与 LDA-AR 模型主题的 KL 距离最小值(即每列的最小值),将其与所有主题 KL 距离的平均值(7.7)进行比对,如果小于平均值,则将其对应。例如,LDA(AR)模型中 Topic 1 与 ATT 模型中 Topic 26 的 KL 距离最小为 4.0(小于 7.7),于是把 LDA(AR)模型中的 Topic 1 与 ATT 模型中的 Topic 26 对应起来。

(2)如果 ATT 模型主题与 LDA-AR 模型主题间的 KL 距离的最小值大于所有主题 KL 距离的平均值(7.7),则 LDA(AR)模型主题与 ATT 模型主题不进行对应,例如,LDA(AR)模型中 Topic 3、Topic 13 与 ATT 模型主题的 KL 距离的最小值分别为 9.1、8.8,则这 2 个主题单独形成主题。

(3)如果 LDA(AR)模型中的 2 个主题的 KL 距离最小值同时对应 ATT 模型中的一个主题,即列的最小值出现在同一行,则取二者中较小值的对应主题。例如,LDA(AR)模型中的 Topic 7、11 的 KL 距离的最小值分别为 5.8、5.5,同时对应 ATT 模型中的 Topic 15,这种情况下取 KL 距离为 5.5 的 Topic 11 与之相对应,而 Topic 7 不与 ATT 模型中的主题对应,作为独立的主题进行后续的观察分析。

按照上述规则,将其他的主题一一对应起来,在本次 2 类模型各选取的 30 个模型中,共有 24 对主题存

在对齐关系,每类模型各剩余 6 个主题没有对齐。为了避免主题间差异减小或者主题缺失等问题,未对齐的 12 个主题各自成为一个主题,其他对齐的 24 个主题成对合并,共形成 36 个主题。在此基础上对作者的研究兴趣进行动态建模。

表 4 所列的是部分对齐主题结果中出现概率最大的 5 个词。分析表 4 可以发现,实验的数据集上对应的主题在词汇分布上具有一定的共现性,例如,LDA-AR 的 Topic 1 和 ATT 中的 Topic 26 均包含“银行”和“服务”, LDA-AR 中的 Topic 2 和 ATT 中的 Topic 24 中均包含“风险”和“管理”,呈现较高的一致性;而在 LDA-AR 的 Topic 1—Topic 5 中的前 5 个代表词中并没有出现相同的或者语义相近的词汇,这表明不同主题之间存在较大差异。

表 4 主题中出现概率最大的词(部分)

Table 4 The words with the highest probability of appearance in each topic(part)

LDA-AR		ATT	
Topic	Words(前 5 个词项)	Topic	Words(前 5 个词项)
Topic 1	银行 0.288 1 商业 0.114 2 服务 0.088 3 机构 0.041 6 票据 0.028 6	Topic 26	银行 0.113 1 服务 0.051 1 风险 0.040 2 直接 0.029 2 中央 0.021 9
Topic 2	风险 0.281 2 管理 0.109 2 巴塞尔协议 0.062 8 金融 0.043 7 操作 0.038 3	Topic 24	市场 0.02 1 风险 0.112 3 管理 0.095 2 业务 0.009 6 分业 0.001 9
Topic 3	金融 0.184 6 汇率 0.172 6 经济 0.154 6 人民币 0.099 6 机构 0.092 3	—	—
Topic 4	代理 0.053 9 现金 0.053 9 模拟 0.024 9 马尔可夫链 0.012 5 委托 0.012 5	Topic 14	外汇 0.064 6 企业 0.021 6 发行 0.021 5 美国 0.008 1 结构化 0.008 1
Topic 5	股票 0.107 8 分行 0.075 7 动态 0.068 2 发展 0.063 1 浦东 0.050 5	Topic 5	波动 0.059 0 分析 0.040 6 收益率 0.036 9 收益 0.036 9 预测 0.033 2

3.4 模型对比分析

为了分析 LDA 模型在加入作者排名特征后的主题强度变化,先获取 LDA 模型主题再离散到时间窗与将时间作为内在变量的模型的主题强度变化,以及 ATT+LDA(AR)复合主题模型的主题强度变化。本文利用传统 LDA 模型、LDA(AR)模型、ATT 模型以及复合演化模型绘制作者巴曙松 Topic 1 的主题强度变化的趋势图,如图 5 所示。

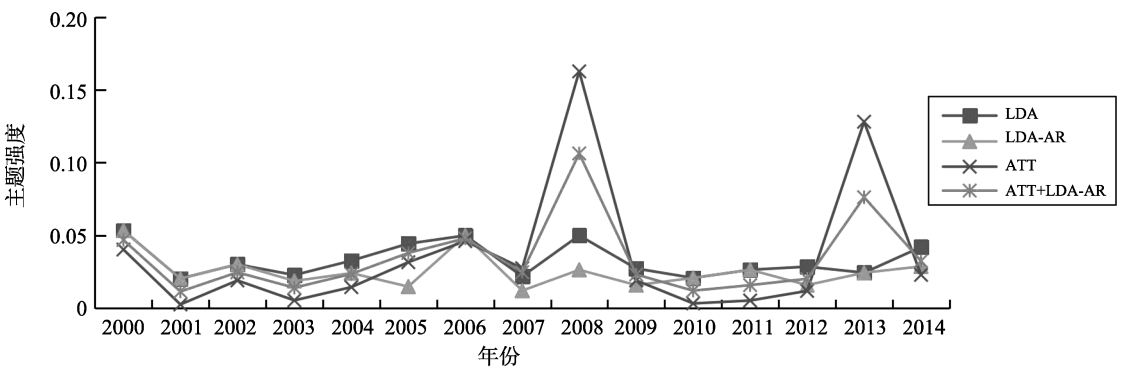

图 5 作者巴曙松 Topic 1 主题趋势变化

Fig.5 Topic 1 trend figure of author BA Shushong

从图 5 可以看出,4 种模型的结果总体趋势呈现一致性,均显示作者对该主题的兴趣在 2000、2006、2008、2013 年左右达到峰值,在 2001、2003、2010 年左右达到低谷。

从局部来看,个体模型的结果存在部分差异性。例如,从 LDA 模型与 LDA(AR)模型的曲线对比来看,2004—2005 年作者巴曙松对于 Topic 1 的兴趣在 LDA 模型中显示为上升趋势,而加入作者排名特征后显示为下降趋势。通过对原始数据进行分析发现,在 2005 年该作者发表的与 Topic 1 相关的文献中,其排名均在较为靠后的位置,这表明 LDA(AR)模型相较于传统 LDA 模型而言,与现实能够更好地拟合。

从 LDA(AR)模型与 ATT 模型的曲线对比来看,二者在 2008、2013 年存在明显的差异。对比二者在这 2 年的兴趣强度可发现,ATT 模型在峰值上远远高于 LDA(AR)模型。这表明将时间信息作为内在变量的

ATT 模型相较于后验的 LDA(AR)模型来说,对于反应同一时间段内主题强度变化更为敏锐。

从 ATT+LDA(AR)复合演化模型与 ATT 模型、LDA(AR)模型的曲线对比来看,对于主题强度的变化幅度相对平缓。可以看出,复合演化模型在 2008、2013 年的峰值数据介于 ATT 模型和 LDA(AR)模型之间。这表明,综合考虑排名特征以及时间因素影响,使将复合演化模型能够有效缓解 ATT 模型对于主题强度变化的敏感性。

鉴于复合演化模型对于获取作者主题强度变化趋势的有效性, 以下将阐述通过复合演化模型所获得的作者兴趣动态变化趋势。

3.5　作者研究兴趣动态变化的结果分析

3.5.1　主题下作者兴趣变化分析

为了分析作者在时间变化下动态的研究兴趣变化,本文选取了知名作者巴曙松、冯宗宪和何德旭等人,利用复合主题演化模型,绘制从 2000—2014 年其对于 Topic 2 的主题强度变化趋势,如图 6 所示。

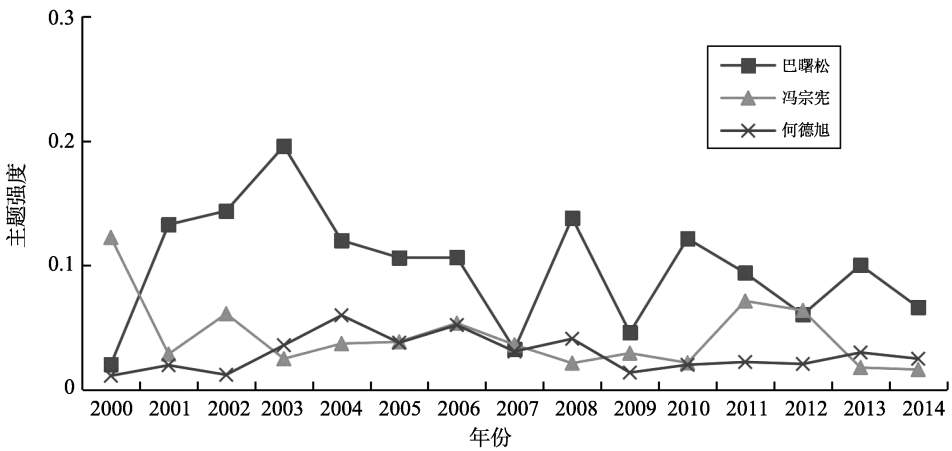

图 6　作者 Topic 2 主题趋势变化

Fig.6　Topic 2 trend figure of authors

从图 6 中的 3 条曲线可以观察出这 3 名作者对于 Topic 2 的研究兴趣变化。其中,作者何德旭和冯宗宪对 Topic 2 的兴趣保持在 0~0.05 之间,这表明这 2 名作者对该主题的兴趣相对较弱;作者巴曙松对 Topic 2 在多个时间段内(2001—2006、2008、2010 年)高于 0.1,表明对该主题保持了较强的研究兴趣。

3.5.2　作者下主题强度变化分析

作者巴曙松在 2000—2014 的时间共发表了 150 篇期刊文献,其中有部分的文献为与其他作者共同发表,考虑到合作作者文献中,作者的排名对于作者主题兴趣的影响,ATT+LDA(AR)模型的作者-主题概率变化趋势如图 7 所示。

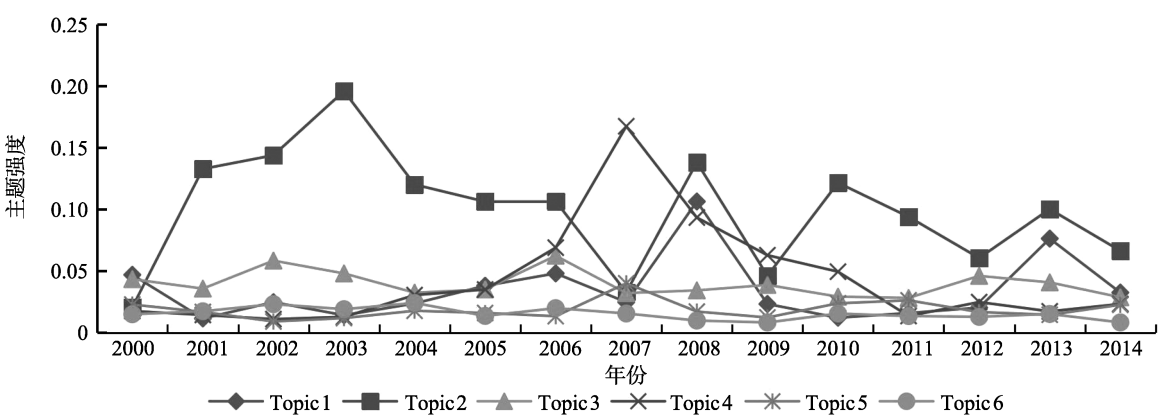

图 7　作者巴曙松部分主题变化

Fig.7　Topic trend figure of author Ba Shushong

图 7 显示了作者对于 6 个主题(Topic 1—Topic 6)的兴趣演化情况。可以看出,在上述 6 个主题中,作者对 Topic 2 表现出最强的兴趣,其主题强度在 2001—2006、2008、2010 年等多个时间段内超过 0.1;其次为

Topic 4,其主题强度在 2007 年超过 0.1;接下来为 Topic 1 和 Topic 3,前者主题强度在 2008 年超过 0.1,后者在 2002、2006 年超过 0.05。

图 8 是作者巴曙松在 2014 年的主题分布图,从图中可以看出在 2014 年该作者最感兴趣的主题是 Topic 30,其次是 Topic 32、Topic 2 和 Topic 11。

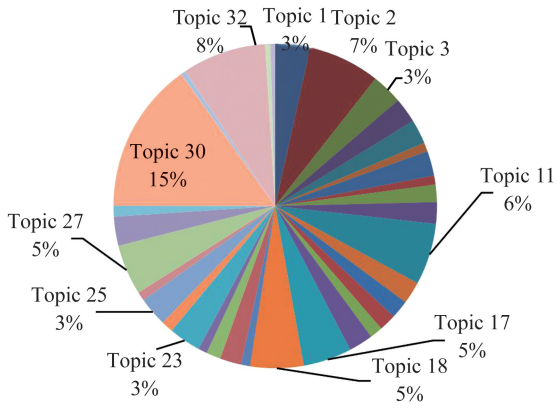

图 8 作者巴曙松 2014 年主题分布图

Fig.8 Topic distribution figure of author BA Shushing in 2014

图 9 是上述 4 个主题(Topic 2、Topic 11、Topic 30 和 Topic 32)的主题河图。图中可以直观地看出多个主题的变化趋势,从 2000—2014 年这 15 年间作者对于这 4 个主题的兴趣保持在较高的水平。在 2003 年,作者对于 4 个主题的兴趣都处在峰值状态;到 2005 年,该作者对于 Topic 2、Topic 6 和 Topic 30 主题的兴趣下滑,对 Topic 32 的兴趣到达另一个峰值(2005 年 Topic 32 的主题强度高于 2003 年)。这表明,作者研究兴趣在 2003—2005 年间更多地转移到 Topic 32 上,从而分散了对其他主题的关注度。

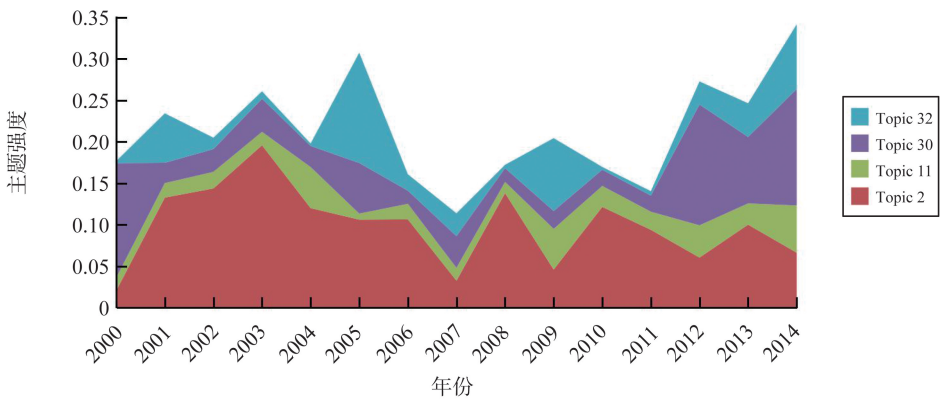

图 9 作者巴曙松主题河图(Topic 2、11、30、32)

Fig.9 Theme river figure of author BA Shushing(Topic 2、11、30、32)

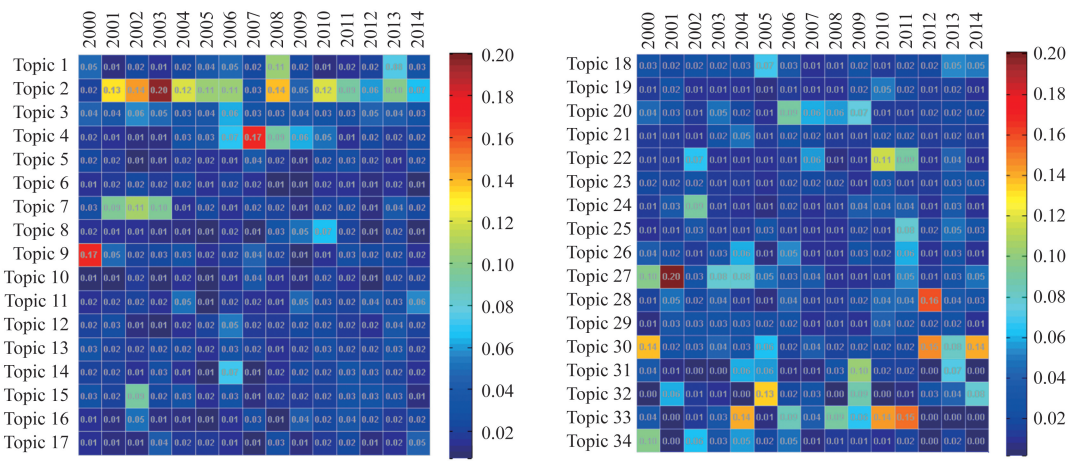

图 10 作者巴曙松主题色块图

Fig.10 Topic color figure of author BA Shushing

图 10 是作者巴曙松在 2000—2014 年 34 个主题的色块图,在图中正方形的色块颜色越接近深蓝色则表示该主题的强度越小,越接近深红色则表示该主题的强度越大;其中,从 15 年的时间跨度上来看,作者巴曙松对与主题 2、4、30、32 这 4 个主题长时间的保持感兴趣的状态,从表 4 中可以看到,这 4 个主题主要是关于“风险管理”“代理”“资本项目”“国际金融”等内容。

本文目的是提出一种复合型主题演化模型,能够有效地揭示金融领域作者研究兴趣的动态变化,由于研

究人员更希望了解在该领域下的资深学者的主题变化,因此本文选取 3 位作者(综合考虑作者的发文量、引用量以及数据库中收录文章的下载量排名前三的作者)的主题演化分析。事实上,由于本文的模型并未对论文作者的输入作相应的限制,因此模型可以应用到数据集中的任意一位作者。

3.5.3　各作者类别主题兴趣分析

为了分析验证在同一时间段下,高产、中产、低产作者感兴趣的主题数量的分布情况。本文将 2014 年的作者按照发表论文数目的多少划分为 4 个组,分别为发表论文数大于 9 篇、发表论文数在 5~9 篇、发表论文数在 3~4 篇和发表论文数在 1~2 篇。其中发表论文数大于 9 篇这一组作为高产作者组,发表论文数在 5~9 篇、发表论文数在 3~4 篇为中产作者组,发表论文数在 1~2 篇为低产作者组。另外定义作者下主题概率超过 0.08 的主题判定为该作者感兴趣的主题,以此来计算作者感兴趣的主题数目。2014 年这 4 个作者组的兴趣主题数分布如表 5 所示。

表 5　2014 年作者兴趣主题数分布

Table 5　The topic number distribution of authors in 2014

作者论文篇数	感兴趣主题数目			
	0	1	2	3
>9	0	0	1(14%)	6(86%)
[5,9]	0	8(7%)	36(33%)	64(60%)
[3,4]	1(0.2%)	93(2%)	183(39.4%)	188(40.4%)
[1,2]	1 508(21%)	3 015(41%)	1 797(25%)	1 012(13%)

表 5 中,中产或者低产作者感兴趣的主题数目为 0 的情况有 2 种:第 1 种是该作者所发表的文档下主题出现概率均小于设置的阈值 0.08;第 2 种情况是该作者发表的文档为合作作者发表的,并且该作者的排名靠后,在 ATT+LDA(AR)模型下削弱了该作者对于文档中主题的兴趣。另外,从表中可以观察出高产作者感兴趣的主题数目为 3 所占的比例为 86%,其次是 2 名中产作者的组所占比例分别为 60%和 40.4%,最后为低产作者组为 13%。感兴趣主题数目为 0 和 1 的表现完全相反,低产作者组占比最高,中产组其次,高产组最低。由此可以了解到在同一时间段下,高产作者感兴趣的主题数目比中产或者低产作者多的可能性很大。这种情况可能是高产作者更专注与该领域的研究,尝试从各个方面研究问题等原因导致的。

4　结语

本文利用主题模型等方法进行主题发现等工作,通过对这些主题信息进行分析处理,能够给寻找科研合作伙伴或者需要了解某个作者的研究动态的用户提供帮助,为寻找合适的项目合作者的科研人员给予推荐意见。本文的创新之处在于:(1)考虑合作作者排名因素。当前对于作者主题的研究中大都没有考虑到合作作者文献中作者排名的特征,然而合作作者排名的因素对于作者兴趣的展现有较大的影响,本文通过在主题模型中加入合作作者排名特征,使得本文提出的模型能够更准确地把握作者的主题兴趣;(2)提出的复合型主题模型结合了先获取主题再离散到时间窗和将时间因素作为主题模型的内在变量 2 种模型,既考虑了时间因素作为内在因子对于主题变化的影响,同时把握了数据集整体的主题内容。

本文的研究存在以下不足:(1)在结合 LDA-AR 模型与 ATT 模型时,对 2 个模型的权重直接取均值,这样虽然在一定程度上简化了算法结构,但是没有考虑权重的设置对作者主题强度变化影响的问题,在后续的研究工作中,将会对实验中的权重参数进行设置调整分析实验的效果;(2)在评估复合主题演化模型对于发现用户的动态兴趣变化的效果方面,本文也存在一些局限,在后续实验中,将采用用户访谈结合人工标注的形式来进一步检验模型之间的差异性;(3)时间切片的大小选择对于主题演化趋势的分析具有较大影响。通常情况下,时间切片过大容易导致主题演化的揭示存在失真现象, 过小则由于引入较多的时间节点,增加了计算的复杂性。本文考虑了语料集合的内容及时间属性,采取了主题演化趋势之中较为常见的时间切片单位(即以年为单位)。在后续研究中,将进一步研究时间切片增加和缩短对用户兴趣演化趋势分析的影响。

参考文献:

[1] BLEI D, NG A, JORDAN M. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[2] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The author-topic model for authors and documents[C]// Conference on Uncertainty in Artificial Intelligence. [S.l.]: AUAI Press, 2004:487-494.

[3] ROSEN-ZVI M, CHEMUDUGUNTA C, GRIFFITHS T, et al. Learning author-topic models from text corpora [J]. ACM Transactions on Information Systems, 2010, 28(1):1-33.

[4] MIMNO D, McCallum A. Expertise modeling for matching papers with reviewers[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007:500-509.

[5] KAWAMAE N. Author interest topic model[C] // Proceedings of the 33th International ACM STGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2010:887-888.

[6] KAWAMAE N. Latent interest-topic model: finding the causal relationships behind dyadic data[C]// Proceedings of the 19th ACM CIKM International Conference on Information and Knowledge Management. New York: ACM Press, 2010:649-658.

[7] 李纲, 李岚凤, 毛进, 等. 作者合著网络中研究兴趣相似性实证研究[J]. 图书情报工作, 2015(2):75-81.
LI Gang, LI Lanfeng, MAO Jin, et al. Empirical research on similarity of research interests in co-authorship network [J]. Library and Information Service, 2015(2):75-81.

[8] 徐戈, 王厚峰.自然语言处理中主题模型的发展[J].计算机学报, 2011, 34(8): 1424-1437.
XU Ge, WANG Houfeng. The development of topic models in natural language processing [J]. Chinese Journal of Computers, 2011, 34(8):1424-1437.

[9] 刘萍, 郑凯伦, 邹德安. 基于LDA模型的科研合作推荐研究[J]. 情报理论与实践, 2015, 38(9):79-85.
LIU Ping, ZHEN Kailun, ZHOU Dean. The scientific collaboration recommendation based LDA model [J]. Information Studies: Theory & Application, 2015, 38(9):79-85.

[10] HAN J, LEE H. Characterizing the interests of social media users: refinement of a topic model for incorporating heterogeneous media [J]. Information Sciences, 2016, 358-359: 112-128.

[11] 余传明, 张小青, 陈雷.基于LDA模型的评论热点挖掘:原理与实现[J].情报理论与实践, 2010(5):103-106.
YU Chuanming, ZHANG Xiaoqing, CHEN Lei. Mining hot topics of user comment based on LDA model: principle & approch[J]. Information Studies: Theory & Application, 2010(5):103-106.

[12] 陈文涛, 张小明, 李舟军.构建微博用户兴趣模型的主题模型的分析[J].计算机科学, 2013, 40(4):127-130.
CHEN Wentao, ZHANG Xiaoming, LI Zhoujun. Analysis of topic models on modeling micro blog user interestingness [J]. Computer Science, 2013, 40(4):127-130.

[13] 单斌, 李芳.基于LDA话题演化研究方法综述[J].中文信息学报, 2010, 24(6):43-49.
SHAN Bin, LI Fang. A survey of topic evolution based on LDA[J]. Journal of Chinese Information Processing, 2010, 24(6):43-49.

[14] BLEI D M, LAFFERTY J D. Dynamic topic models[C]// Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006: 113-120.

[15] WANG X, MCCALLUM A.Topics over time:a non-Markov continuous-time model of topical trends[C]// Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining.New York: ACM, 2006:424-433.

[16] XU Shuo, SHI Qingwei, QIAO Xiaodong, et al. Author-topic over time (AToT): a dynamic users' interest model [J]. Lecture Notes in Electrical Engineering, 2014, 274:58-61.

[17] 倪丽萍, 刘小军, 马驰宇, 基于LDA模型和AP聚类的主题演化分析[J].计算机技术与发展, 2016, 26(12):6-11.
NI liping, LIU Xiaojun, MA Chiyu. Topic evolution analysis based on LDA model and AP clustering [J]. Computer Technology and Development, 2016, 26(12):6-11.

[18] 史庆伟, 李艳妮, 郭朋亮, 科技文献中作者研究兴趣动态发现[J].计算机应用, 2013, 33(11):3080-3083.
SHI Qingwei, LI Yanni, GUO Pengliang. Dynamic finding of authors' research interests in scientific literature [J]. Journal of Computer Applications, 2013, 33(11):3080-3083.

[19] 廖君华, 孙克迎, 钟丽霞. 一种基于时序主题模型的网络热点话题演化分析系统[J].图书情报工作, 2013,57(9):96-102.
LIAO Junhua, SUN Keying, ZHONG Lixia. Study a hot topics analysis system based on time sliced topic model [J]. Library and Information Service, 2013, 57(9):96-102.

[20] 李保利, 杨星. 基于LDA模型和话题过滤的研究主题演化分析[J]. 小型微型计算机系统, 2012(12): 2738-2743.
LI Baoli, YANG Xing. Analyzing research topic evolution with LDA and topic filtering[J]. Journal of Chinese Computer Systems, 2012(12): 2738-2743.

[21] YAN Erjia. Research dynamics, impact, and dissemination: a topic-level analysis[J]. Journal of the Association for Information Science & Technology, 2015, 66(11):2357-2372.

(编辑:许力琴)