情 报 学 报 ISSN 1000 - 0135
第 32 卷 第 9 期 912 - 919, 2013 年 9 月
JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 32 No. 9, 912 - 919 September 2013

doi:10.3772/j.issn.1000-0135.2013.09.002

作者主题演化模型及其在研究兴趣演化分析中的应用[1)]

史庆伟[1,2] 乔晓东[1] 徐 硕[1] 农国武[1,3]

(1. 中国科学技术信息研究所信息技术支持中心,北京 100038;
2. 辽宁工程技术大学软件学院,葫芦岛 125105;3. 中国铝业广西分公司信息部,平果 531400)

摘要 从海量科技文献中自动挖掘隐含主题、研究人员的研究兴趣及其演化规律是信息服务迈向知识服务需要解决的关键问题之一。目前的方法多从静态的角度分析文献主题、科研人员的研究兴趣,而演化分析的方法主要集中文档的内部特征,即文档内容本身,很少考虑作者等外部特征。基于此,本文在 AT 和 ToT 模型的基础上构建了作者主题演化(AToT)模型,并给出了一种估计 AToT 模型参数的吉布斯采样方法。该模型集成了 AT 和 ToT 模型的优势,不仅可以揭示科技文献中隐含的主题、作者的研究兴趣,而且可以挖掘研究兴趣随时间变化的规律。最后,以 1740 篇 NIPS 会议论文集作为实验数据,通过与 AT 模型的对比分析验证了 AToT 模型的可行性和有效性。

关键词 主题模型 作者主题演化模型 研究兴趣演化分析 吉布斯采样 困惑度

Author-Topic Evolution Model and Its Application in Analysis of Research Interests Evolution

Shi Qingwei[1,2], Qiao Xiaodong[1], Xu Shuo[1] and Nong Guowu[1,3]

(1. *Information Technology Supporting Center, Institute of Scientific and Technical Information of China, Beijing* 100038;
2. *School of Software, Liaoning Technical University, Huludao* 125105;
3. *Information Department, CHALCO Guangxi Branch, Pingguo* 531400)

Abstract One of the key problems in upgrading information services towards knowledge services is to automatically mine latent topics, researchers' interests and their evolution patterns from large-scale scientific & technical literatures. Most of current methods are devoted to discover static latent topics and research interests. Nevertheless, previous evolution analysis research mainly focuses on analyzing intra-features of documents, namely documents' text content without considering directly extra-features of documents such as authors. To overcome this problem, on the basis of Author-Topic (AT) model and Topics over Time (ToT) model, Author-Topic over Time (AToT) model is constructed in this study, and Gibbs sampling method is utilized to estimate corresponding parameters. This model is not only able to discover latent topics and researchers' interests, but also to mine their changing patterns over time. Another way to say this is that our AToT model combines the advantages of AT and ToT models. Finally, the extensive experimental results on NIPS dataset with 1740 papers indicate that our AToT model is feasible and efficient.

Keywords topic model, author-topic (AT) model, research interests analysis, gibbs sampling, perplexity

收稿日期:2012 年 12 月 1 日

作者简介:**史庆伟**,男,1973 年生,博士,副教授,主要研究方向:信息检索、文本挖掘和机器学习等。E-mail: shiqw@istic.ac.cn。**乔晓东**,男,1965 年生,英国谢菲尔德大学硕士,研究员,主要研究方向:信息服务、信息资源管理等。**徐硕**(通讯作者),男,1979 年生,博士,副研究员,主要研究方向:数据挖掘、机器学习和知识工程等。E-mail: xush@istic.ac.cn。**农国武**,男,1970 年生,硕士,高级工程师,主要研究方向:信息服务、知识工程等。

1) 本研究受"十二五"国家科技支撑计划"面向外文科技知识组织体系的大规模语义计算关键技术研究"(2011BAH10B04)、"基于 STKOS 的知识服务应用示范"(2011BAH10B06)以及中国科学技术信息研究所预研项目"基于词系统的领域深层主题规律揭示分析研究"(YY201216)资助。

1 引 言

科技文献作为学术成果的主要载体,凝聚了人类的大量智慧,是传播知识、进行学术交流的窗口。牛顿那句为人熟知的名言:“如果我看得更远的话,那是因为我站在巨人的肩膀上”很好地诠释了任何科学研究的成果都是建立在前人基础之上的。然而,科学知识/主题的演化规律往往不是显而易见的,一般也只有极少数的领域专家可以谙熟于心。

普赖斯科技文献指数增长定律[1]和逻辑曲线增长模型[2]表明,科技文献量在快速增长,这给科学知识/主题的人工探测与跟踪带来了巨大的挑战。为了准确把握科技发展现状,如何结合科技文献的内外部特征从海量科技文献中自动发现科技主题及其内部的发展脉络成为目前亟待解决的关键问题之一。

科技文献资源包含大量的隐含信息,如词与词之间的潜在语义关系、文献主题与作者的关系(作者的研究兴趣)和研究热点的兴起、成熟到逐渐衰退的过程等。传统的信息分析方法难于捕获这些潜在信息,因此无法满足用户对科技信息深层次的需求。近年来,以 Blei 等提出的 LDA(Latent Dirichlet Allocation)模型[3]为代表的产生式模型在表示文档、模拟文档的产生过程、处理文档降维、挖掘文档中隐含信息等方面取得了长足进步,已经被广泛应用于信息抽取、社会媒体挖掘和学术挖掘等领域。

在科研人员研究兴趣挖掘方面,Rosen-Zvi 等在 LDA 模型中引入作者隐变量,用作者 - 主题分布取代 LDA 模型中文档 - 主题分布,提出了 AT(Author-Topic)模型[4,5]。该模型可以有效地挖掘作者与主题之间的联系,即科研人员的研究兴趣。然而,该模型隐式地假设每个科研人员只有一个研究兴趣,这有悖于实际情况。为克服这一限制,Mimno 等在 AT 模型的基础上构建了 APT(Author-Persona-Topic)模型[6]。该模型将“身份”(Persona)与研究兴趣相对应,并给出了一种估计研究兴趣个数的启发式方法。实际上,AT 和 APT 模型在挖掘科研人员的研究兴趣时,只考虑了其撰写的文献(无论第几作者),而丢弃了与其研究兴趣类似的其他科研人员所撰写的文献。换句话说,AT 和 APT 模型是在局部而非全局信息的基础上挖掘科研人员的研究兴趣。Kawamae 提出的 AIT(Author-Interest-Topic)模型[7,8],通过“文档类”的概念放宽了这种限制。

其实,在整个科研生涯过程中,每个科研人员的研究兴趣通常不是一成不变的。比如,著名的科学家牛顿在力学、光学和微积分等领域都有卓越的贡献,但他早期的研究兴趣多与经典力学有关,后期发表的作品集中微积分和光学领域。AT、APT 和 AIT 模型在建模研究兴趣时,均未直接考虑时间因素的影响。当然,可以通过预处理[9]或后处理[5]的方式引入时间因素,但这些方法容易导致局部主题的淹没[10]或主题对应困难[11]等问题。

直到 2006 年,Blei 等借助时间序列分析方法构建了 DTM(Dynamic Topic Model)模型[12],才真正将时间因素有机集成到主题模型中。DTM 将整个文档集划分到不同的时间窗口中,利用 LDA 模型对每个窗口内的文档子集进行建模分析,为降低计算复杂度,假设当前时间窗口的模型参数仅与前一时间窗口的模型参数有关,即不同时间窗口的模型参数服从一阶马尔可夫假设,最后利用状态空间模型实现主题演化分析。不过,由于 DTM 模型对时间的离散化处理,使得该模型的实际效果对时间粒度特别敏感。为此,Wang 等利用布朗运动模型将文本的时间戳信息引入到参数演化过程中,构建了连续时间版本的 DTM 模型(cDTM)模型[13]。然而,大量研究表明[14~16],主题的演化过程经常呈现跳跃性,也就是说主题演化并不一定服从一阶马尔可夫假设。考虑到贝塔分布密度函数的形状比高斯分布的更丰富,Wang 等假设时间服从贝塔分布,提出了与马尔可夫假设无关的 ToT(Topic over Time)模型[17]。

从以上分析不难看出,AT、APT 以及 AIT 等模型主要从静态的角度分析科研人员的研究兴趣,尽管可以通过预处理或后处理的方式引入时间因素,而演化分析主要集中文档的内部特征,即文档内容本身,很少考虑作者等外部特征。基于此,本文在 AT 和 ToT 模型的基础上,提出了一种作者主题演化(Author-Topic over Time, AToT)模型,该模型不仅可以分析科研人员与主题的关系,而且可以揭示研究人员研究兴趣随时间的变化规律,如图 1 所示。每个作者对应一个主题概率分布;每个主题对应一个词项概率分布和一个随时间变化的贝塔分布。需要说明的是,本文所述方法同样可以将 ToT 模型与 APT、AIT 等静态研究兴趣分析模型相结合构建新的作者主题演化模型,为简单起见,本文仅以 AT 模型为例。

本文其余部分的组织结构如下:第 2 节详细介

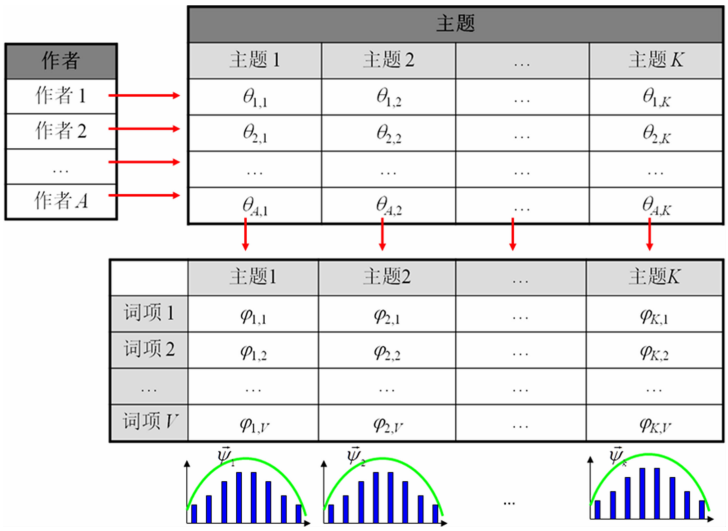


图 1　作者、主题和时间的关系

绍 AToT 模型构建方法；第 3 节介绍利用吉布斯采样方法估计 AToT 模型中的参数；第 4 节为实验结果分析及评价部分；第 5 节总结全文。

2　作者主题演化(AToT)模型

为叙述方便，表 1 集中对本文所使用的符号进行说明。

表 1　符号说明

符号	描述
K	主题的数量
M	文档的数量
V	词项的数量
A	作者的数量
N_m	文档 m 中单词的数量，即文档 m 的长度
A_m	撰写文档 m 的作者数量
$\mathbf{a}_m$	撰写文档 m 的作者形成的向量
ϑ_a	作者 a 的主题概率分布
φ_k	主题 k 的词项概率分布
ψ_k	主题 k 随时间变化的贝塔分布
$z_{m,n}$	文档 m 中第 n 个单词的主题分配
$x_{m,n}$	文档 m 中第 n 个单词的作者分配
$w_{m,n}$	文档 m 中第 n 个单词
$t_{m,n}$	文档 m 中第 n 个单词的时间戳
α	$\alpha_a(a = 1, \cdots, A)$的超参数
β	$\beta_k(k = 1, \cdots, K)$的超参数

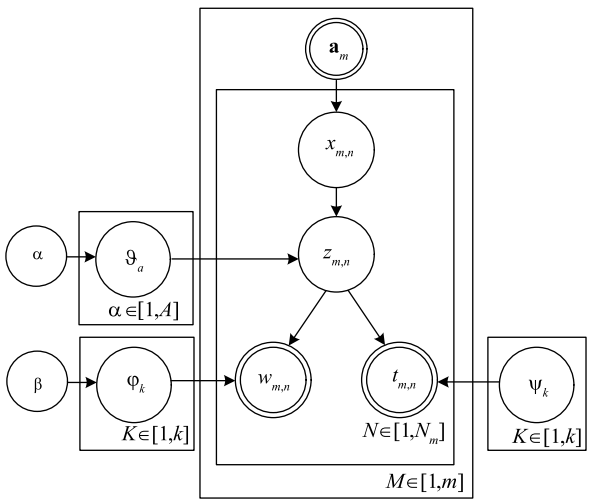


图 2　AToT 模型概率图

AToT 模型概率图如图 2 所示，该模型的产生过程如下：

(1)对于每个主题 $k \in [1, K]$：

抽取多项式概率分布 $\varphi_k \sim \text{Dirichlet}(\ )$；

(2)对于每个作者 $a \in [1, A]$：

抽取多项式概率分布 $\vartheta_a \sim \text{Dirichlet}(\ )$；

(3)对于每篇文档 $m \in [1, M]$中的每个单词 $n \in [1, N_m]$：

(a)抽取一个作者 $x_{m,n} \sim \text{Uniform}(\mathbf{a}_m)$

(b)抽取一个主题 $z_{m,n} \sim \text{Multinomial}(\vartheta x_{m,n})$

(c)抽取一个单词 $w_{m,n} \sim \text{Multinomial}(\varphi z_{m,n})$

(d)抽取一个时间戳 $t_{m,n} \sim \text{Beta}(\psi z_{m,n})$

需要说明的是，单词和时间戳的抽取无先后顺序之分。

3 参数估计

ATOT 模型有三个未知参数集 $\Phi=\{\varphi_k\}K\ k=1$、$\Theta=\{\vartheta_a\}A\ a=1$ 和 $\Psi=\{\Psi_k\}K\ k=1$ 需要估计。目前估计这些参数的方法有很多，如变分期望最大化(Variational Expectation Maximization)[3]、期望传播(Expectation Propagation)[18,19]以及 Collapsed 吉布斯采样[20]等。每种参数估计方法都各有利弊，选择一种合适的近似算法要在效率、复杂性、准确性和概念简洁性之间综合考虑[21]。由于 Collapsed 吉布斯采样方法描述简单且更易于实现，成为主题模型中最常采用的参数估计方法，也是本文所采用的参数估计方法。

所谓“Collapsed”，是指通过积分巧妙地避开了实际待估计的参数 $\Phi=\{\varphi_k\}K\ k=1$ 和 $\Theta=\{\vartheta_a\}A\ a=1$ 转而对每个单词 $w_{m,n}$ 的主题 $z_{m,n}$ 和作者 $x_{m,n}$ 进行采样，一旦每个 $w_{m,n}$ 的 $z_{m,n}$ 和 $x_{m,n}$ 确定下来，Φ 和 Θ 的值可以很容易通过统计频次后计算出来。

吉布斯采样是 MCMC(Markov-Chain Monte Carlo)[22]的特例，每次对联合分布的一个分量进行采样，而保持其他分量的值不变。经推导，ATOT 模型的吉布斯采样公式(该公式通常被称为全条件概率(Full Conditionals))为：

$$P(z_{m,n},x_{m,n}\mid w,z_{\neg(m,n)},x_{\neg(m,n)},t,a,\alpha,\beta,\Psi)\propto\frac{n_{z_{m,n}}^{(w_{m,n})}+\beta_{w_{m,n}}-1}{\sum_{v=1}^{V}(n_{z_{m,n}}^{(v)}+\beta_v)-1}\times\frac{n_{x_{m,n}}^{(z_{m,n})}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n_{x_{m,n}}^{(k)}+\alpha_k)-1}\times\frac{\Gamma(\psi_{z_{m,n},1}+\psi_{z_{m,n},2})}{\Gamma(\psi_{z_{m,n},1})\Gamma(\psi_{z_{m,n},2})}t_{m,n}^{\psi_{z_{m,n},1}-1}(1-t_{m,n})^{\psi_{z_{m,n},2}-1}\quad(1)$$

其中，$z_{\neg(m,n)},x_{\neg(m,n)}$ 表示除了分配到单词 $w_{m,n}$ 的主题和作者以外的所有主题、作者变量，$n(v)$ $z_{m,n}$ 表示词项 v 被分配主题 $z_{m,n}$ 的次数，$n(k)$ $x_{m,n}$ 表示作者 $x_{m,n}$ 所负责的单词被分配主题 k 的次数。

一旦通过吉布斯采样为每个单词 $w_{m,n}$ 分配了主题 $z_{m,n}$ 和作者 $x_{m,n}$，利用 Dirichlet 分布的期望，可很容易估算参数 φ_k 和 ϑ_a：

$$\varphi_{k,v}=\frac{n_k^{(v)}+\beta_v}{\sum_{v=1}^{V}(n_k^{(v)}+\beta_v)}\quad(2)$$

$$\vartheta_{a,k}=\frac{n_a^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_a^{(k)}+\alpha_k)}\quad(3)$$

单词 $w_{m,n}$ 的时间戳对应于文档 m 的时间戳，服从每个主题下的贝塔分布 ψ_k。文献[23]对于 ψ 的估计方法进行了详细介绍，本文出于计算速度考虑，采用矩估计方法(method of moments)估计 ψ_k：

$$\hat{\psi}_{k,1}=t_k\left(\frac{t_k(1-t_k)}{s_k^2}-1\right)\quad(4)$$

$$\hat{\psi}_{k,2}=(1-t_k)\left(\frac{t_k(1-t_k)}{s_k^2}-1\right)\quad(5)$$

其中，t_k 和 s_k^2 分别是主题 k 采样的均值和方差，具体定义如下：

$$t_k=\frac{\sum_{m=1}^{M}(n_m^{(k)}\times t_m)}{\sum_{v=1}^{V}n_k^{(v)}}\quad(6)$$

$$s_k^2=\frac{\sum_{m=1}^{M}(n_m^{(k)}\times t_m^2)}{\sum_{v=1}^{V}n_k^{(v)}}-t_k^2\quad(7)$$

$n_m^{(k)}$ 表示文档 m 中的单词被分配主题 k 的次数。

4 实验结果及分析

4.1 实验数据

NIPS 数据集由 1987 ~ 1999 年 NIPS(Neural Information Processing System)国际会议论文全文文本信息组成，共 1740 篇，各年份的论文量及百分比见表 2。经过去掉停用词、数字和出现次数少于 5 次的单词等预处理工作得到的文本数据包括：13 649个词项，2 301 375 个单词，2037 个作者。

表 2 NIPS 数据集每年论文数量

年份	论文数量(百分比)	年份	论文数量(百分比)
1987	90(5.2%)	1988	95(5.5%)
1989	101(5.8%)	1990	143(8.2%)
1991	144(8.3%)	1992	127(7.3%)
1993	144(8.3%)	1994	140(8.0%)
1995	152(8.7%)	1996	152(8.7%)
1997	151(8.7%)	1998	151(8.7%)
1999	150(8.6%)		

为便于比较，类似于文献[4]和文献[5]，本文将 NIPS 数据进一步分为两部分：1557 篇文档作为训练集，183 篇文档作为测试集，其中测试集中包含 102 篇单作者文档，而且测试集中出现的所有作者必须在训练集中也出现，这样便于分析特定科研人员研究兴趣的泛化能力。

4.2 评价指标

困惑度(perplexity)[24]是评价模型泛化能力的

标准指标,困惑度值越小,说明模型泛化能力越强。ATOT 模型中,对于测试集中的文档 $\tilde{m}$,困惑度计算公式如下:

$$perplexity(w_{\tilde{m},.}, t_{\tilde{m},.} \mid a_{\tilde{m}}, \alpha, \beta, \Psi) = \exp\left[-\frac{\ln P(w_{\tilde{m},.}, t_{\tilde{m},.} \mid a_{\tilde{m}}, \alpha, \beta, \Psi)}{N_{\tilde{m}}}\right] \quad (8)$$

其中,

$$P(w_{\tilde{m},.}, t_{\tilde{m},.} \mid a_{\tilde{m}}, \alpha, \beta, \Psi) = \sum_{z_{\tilde{m},.}} p(t_{\tilde{m},.} \mid \psi_{z_{\tilde{m},n}}) \times \int p(\Phi \mid \beta) \sum_{z_{\tilde{m},.}} \vartheta_{x_{\tilde{m},n}, z_{\tilde{m},n}} d\Phi \times \frac{1}{[A_{\tilde{m}}]^{N_{\tilde{m}}}} \times \int p(\Theta \mid \alpha) \sum_{x_{\tilde{m},.}} \varphi_{z_{\tilde{m},n}, w_{\tilde{m},n}} d\Theta \quad (9)$$

根据训练集得到的参数 Φ、Θ 和 Ψ,利用公式(2)、(3)、(4)和公式(5)可以估计公式(9)中的参数 Φ、Θ 和 Ψ。为尽量准确逼近公式(9)的真实值,本文对测试集运行 $S(=10)$ 次吉布斯采样,困惑度取 S 次采样的平均值。

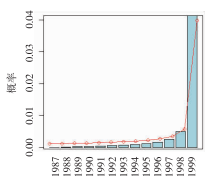
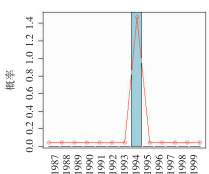
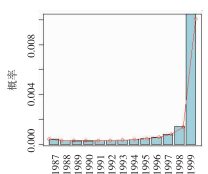
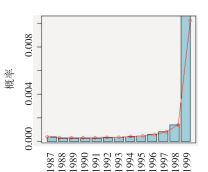
4.3 实验结果分析

本文采用对称 Dirichlet 超参数,即($\alpha_k = 50/K$ ($k \in [1, K]$),($\beta_v = 0.1$ ($v \in [1, V]$),迭代次数设为 2000。

(1)主题揭示分析

类似于文献[4],图 3 给出了利用 ATOT 模型计算得到的 8 个主题。每个主题的描述包括三部分:(a)与主题最相关的前 10 个词项;(b)与主题最相关的前 10 个作者;(c)主题随时间的变化趋势。

主题 87 “支持向量机和核方法”		主题 37 “神经网络”		主题 11 “增强学习”		主题 88 “EM 和混合模型”	
词项	概率	词项	概率	词项	概率	词项	概率
set	0.0188195	learning	0.01106740	state	0.0468466	density	0.0279477
support	0.0187117	network	0.00948016	learning	0.0252876	log	0.0217790
vector	0.0186039	neural	0.00780503	belief	0.0213999	distribution	0.0186946
kernel	0.0160163	input	0.00682192	policy	0.0182191	mixture	0.0178379
function	0.0146146	model	0.00681643	function	0.0175122	method	0.0144108
svm	0.0138060	training	0.00604202	action	0.0150383	gaussian	0.0142394
training	0.0129974	data	0.00597611	states	0.0148615	likelihood	0.0140681
problem	0.0124583	figure	0.00594316	reinforcement	0.0118574	entropy	0.0132113
space	0.0119731	networks	0.00560813	actions	0.0118574	gaussians	0.0123546
solution	0.0115957	function	0.00554222	mdp	0.0102670	form	0.0113264
作者	概率	作者	概率	作者	概率	作者	概率
Scholkopf_B	0.949692	Reggia_J	0.979832	Zhang_N	0.629412	Barron_A	0.608507
Crisp_D	0.888975	Todorov_E	0.976750	Rodriguez_A	0.578235	Wainwright_M	0.372871
Laskov_P	0.706170	Horne_B	0.974146	Dietterich_T	0.342954	Mukherjee_S	0.340927
Steinhage_V	0.634973	Thmn_S	0.973083	Sallans_B	0.228042	Li_J	0.337108
Chapelle_O	0.610385	Weigend_A	0.972806	Walker_M	0.189143	Jebara_T	0.253203
Li_Y	0.513418	McCallum_R	0.969777	Koller_D	0.1885150	Millman_K	0.171569
Herbrich_R	0.454384	Camana_R	0.969388	Yeung_D	0.1213730	Fisher_J	0.148230
Gordon_M	0.425090	Slaney_M	0.969382	Thrun_S	0.0842081	Ihler_A	0.128369
Vapnik_V	0.330421	Miikkulainen_R	0.968541	Konda_V	0.0680365	Beal_M	0.126578
Dom_B	0.286036	Bergen_J	0.968358	Parr_R	0.0468006	Hansen_L	0.0849109

主题 47 “语音识别”		主题 78 “贝叶斯学习”		主题 51 “脸谱识别与因子分析”		主题 58 “数据模型与学习算法”	
词项	概率	词项	概率	词项	概率	词项	概率
hmm	0.0415364	bayesian	0.0243032	sejnowski	0.0265409	learning	0.00904655
speech	0.0392921	sampling	0.018456	eye	0.0265409	model	0.00752741
hmms	0.0216579	prior	0.0178563	ica	0.0183324	neural	0.00705102
mixture	0.0179708	distribution	0.0148578	vor	0.0159531	data	0.00700339
suffix	0.0104362	monte	0.0127588	disparity	0.0153583	function	0.0068393
probabilistic	0.00995527	carlo	0.0118592	head	0.0135738	network	0.00624646
probabilities	0.00947434	model	0.0109597	position	0.0125031	input	0.00593946
singer	0.0088331	posterior	0.0105099	eeg	0.0119083	set	0.00561128
acoustic	0.0088331	priors	0.00946041	parietal	0.0109566	networks	0.00556365
saul	0.00867279	sample	0.00901063	salk	0.0105997	figure	0.00545249
作者	概率	作者	概率	作者	概率	作者	概率
Rigoll_G	0.460882	Schuurmans_D	0.651505	Sejnowski_T	0.410459	Gray_M	0.974482
Singer_Y	0.437547	Sykacek_P	0.495506	Pouget_A	0.269781	Dimitrov_A	0.973538
Nix_D	0.192342	Andrieu_C	0.413324	Anastasio_T	0.112957	Galperin_G	0.97094
Saul_L	0.170699	Rasmussen_C	0.344185	Horiuchi_T	0.0328485	Malik_J	0.968536
Hermansky_H	0.0795602	Zlochin_M	0.244745	Albright_T	0.0099278	Davies_S	0.966534
Roweis_S	0.0391364	Beal_M	0.157807	Jousmaki_V	0.00791139	Cook_G	0.96519
Attias_H	0.0357538	Hansen_L	0.122773	Fredholm_H	0.00681818	Ghosn_J	0.964184
Movellan_J	0.033414	Herbrich_R	0.0882701	Bohr_J	0.00643777	Orponen_P	0.964184
Schuster_M	0.0293324	Downs_O	0.0694726	Ramanujam_N	0.00621891	Yen_S	0.963001
Muller_K	0.028258	Williams_C	0.0652069	Dixon_L	0.00585938	Chatterjee_C	0.962627

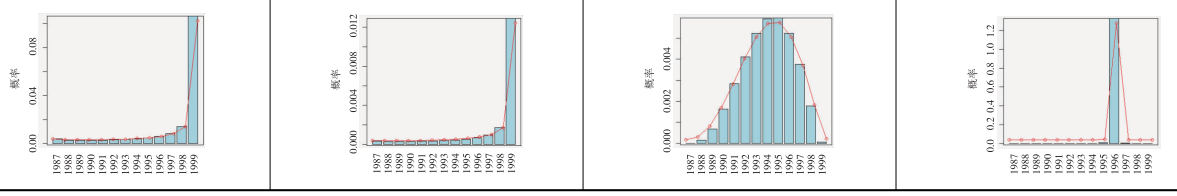


图 3　AToT 模型 NIPS 数据集上计算得到的 8 个主题

从图 3 中可以看出,这 8 个主题分别与“支持向量机和核方法”(主题 87)、“神经网络”(主题 37)、“增强学习”(主题 11)、“EM 和混合模型”(主题 88)、“语音识别”(主题 47)、“贝叶斯学习”(主题 78)、“脸谱识别与因子分析”(主题 51)和“数据模型与学习算法”(主题 58)相关。作者与主题的关系也有很好的描述,如“支持向量机和核方法”主题对应的前四个作者中,Scholkopf 是机器学习领域中核方法的著名专家,通过查找相关作者主页了解到,Crisp、Laskov 和 Steinhage 等作者在支持向量机方面做了大量研究工作。在图 3 中还可以看出,“神经网络”和“数据模型和学习算法”的概率值大说明受到的关注度高。“支持向量机和核方法”、“增强学习”、“EM 和混合模型”、“语音识别”、“贝叶斯学习”等主题在文档集时间后期(1998 年、1999 年)受到的关注度增长很快,而“脸谱识别与因子分析”的关注度在 1994 年、1995 年达到峰值,随后逐渐下降。

(2)研究兴趣演化分析

为了进一步分析科研人员研究兴趣在不同时间阶段的变化情况,需要计算给定作者条件下主题和时间的联合分布,如公式(10)所示:

$$P(z,t \mid a) = P(z \mid a)p(z \mid t) = \vartheta_{a,z} \times \frac{\Gamma(\psi_{z,1}+\psi_{z,2})}{\Gamma(\psi_{z,1})\Gamma(\psi_{z,2})} t^{\psi_{z,1}-1}(1-t)^{\psi_{z,2}-1} \tag{10}$$

下面以研究人员 Sejnowski 为例,说明研究人员研究兴趣的变化规律。Sejnowski 在 1987 ~ 1999 年 NIPS 会议上共发表文献 43 篇,每年的文献数量如

图 4 所示。

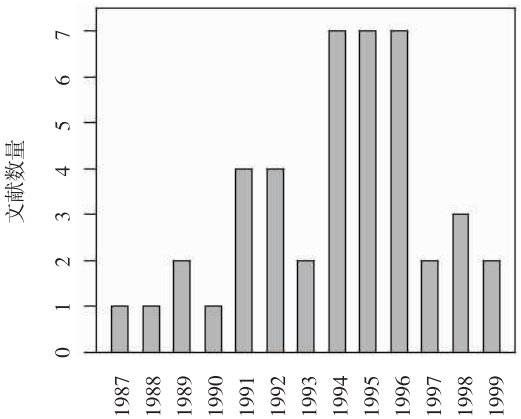


图 4　Sejnowski 在各年发表文献数量分布图

Sejnowski 在不同时期研究兴趣的变化情况，如图 5 所示。图 5 中 Sejnowski 在不同时期对不同主题感兴趣的程度表示为矩形的面积，矩形面积越大，说明研究人员对相应主题感兴趣的程度越高。

从图 5 可以看出，在 1987 ~ 1999 年期间，Sejnowski 的研究兴趣主要是“脸谱识别与因子分析”（主题 51），“神经网络”（主题 37）和“数据模型与学习算法”（主题 58），但不同时间阶段侧重点不同。Sejnowski 早期（1989 ~ 1993 年）的研究兴趣是“脸谱识别与因子分析”，从 1994 年开始 Sejnowski 的研究兴趣扩展到“神经网络”（1994 年）和“数据模型与学习算法”（1996 年），而且研究强度较大（发表文献数量增多）。1997 年以后 Sejnowski 的研究兴趣又回到“脸谱识别与因子分析”上，研究强度也有所下降。

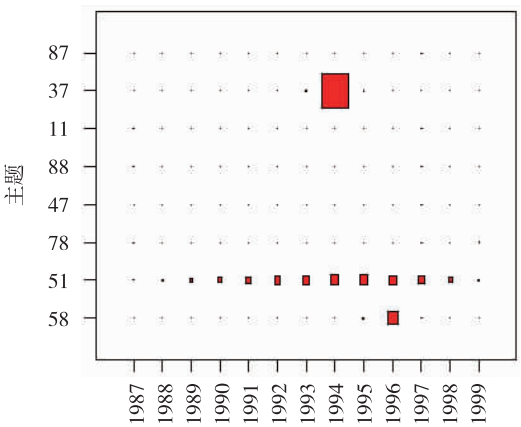


图 5　Sejnowski 的研究兴趣变化示意图

（3）泛化能力分析

为了分析 AToT 模型的泛化能力，本文将 102 篇单作者文档作为测试集，根据公式（9）分别计算了 AT、AToT 模型在不同主题个数情况下相应的困惑度值，实验结果如图 6 所示。从图 6 可以看出，主题数量超过 10 个时，AToT 模型的困惑度明显小于 AT 模型，说明 AToT 模型的性能要优于 AT 模型。

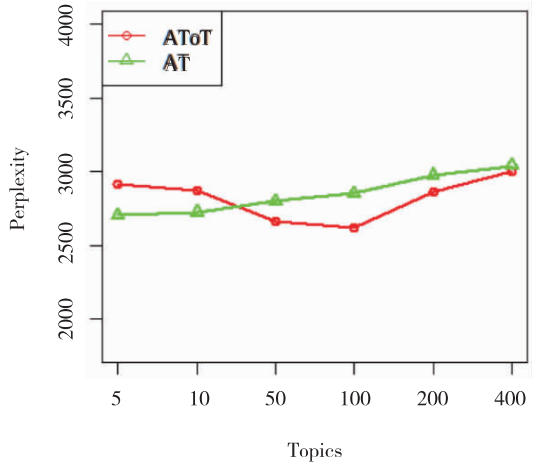


图 6　AToT 与 AT 模型困惑度比较

5　结论及讨论

从海量科技文献中自动挖掘隐含主题、研究人员的研究兴趣及其演化规律是信息服务迈向知识服务需要解决的关键问题之一。目前分析科研人员研究兴趣的方法多从静态的角度考虑文献主题与作者之间的关系，如 AT、APT 和 AIT 模型等。另一方面，文献主题演化分析的方法主要是考虑文档内容本身，如 DTM 和 ToT 模型等，而文档的外部特征如作者等信息则很少被利用。

基于此，本文集成了 AT 和 ToT 模型的优势，构建了作者主题演化（AToT）模型。AToT 模型中，文献隐含主题由主题 - 词项的概率分布描述，主题 - 词项的概率分布由文档中单词共现和文档时间戳决定，从而揭示主题的演化规律。研究人员的研究兴趣由作者 - 主题分布描述，结合主题演化规律可以分析研究人员研究兴趣随时间变化的规律。模型参数的估计通过吉布斯采样的方法实现。最后，以 1740 篇 NIPS 会议论文集作为实验数据，通过与 AT 模型的对比分析验证了 AToT 模型的有效性。

参　考　文　献

[1] Price D S. Little science, big science. New York: Columbia University Press, 1963.

[2] 王崇德. 文献计量学教程[M]. 天津：南开大学出版社，1990.

[3] Blei D, Ng A, Jordan M. Latent Dirichlet Allocation [J]. Journal of Machine Learning Re-search, 2003, 3:

993-1022.

[4] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]// Proceedings of the 20th conference on uncertainty in artificial intelligence (UAI), Arlington: AUAI Press 2004: 487-494.

[5] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery [C]// Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, New York: ACM Press, 2004: 306-315.

[6] 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.

[7] Kawamae N. Author interest topic model[C]// Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, New York: ACM Press, 2010: 887-888.

[8] 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.

[9] Wang X, Mohanty N, McCallum A. Group and topic discovery from relations and text[C]//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM Press, 2005: 28-35.

[10] Song X, Lin C, Tseng B L, et al. Modeling and predicting personal information dissemination behavior [C]// Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM Press, 2005: 479-488.

[11] Xu S, Zhu L, Qiao X, et al. Topic Linkages between Papers and Patents[C] // Proceedings of the 4th AST International Conference on Advanced Science and Technology, Daejeon: SERSC press, 2012: 176-183.

[12] Blei D, Lafferty J. Dynamic Topic Model[C]//Proceedings of the 23rd ICML international conference on Machine learning, New York: ACM Press, 2006: 113-120.

[13] Wang C, Blei D, Heckerman D. Continuous Time Dynamic Topic Models [C]// Proceedings of the 24th conference on Uncertainty in artificial intelligence (UAI), Corvallis: AUAI Press 2008: 579-586.

[14] Wei X, Sun J, Wang X. Dynamic mixture models for multiple time series [C]//Proceedings of the 20th International Joint Conference on Artificial Intelligent, Hyderabad, India: AAAI Press 2007: 2909-2914.

[15] Nallapati R, Ditmore S, Ditmore S, et al. Multiscale topic tomography[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM Press, 2007: 520-529.

[16] Iwata T, Yamada T, Sakurai Y, et al. Sequential Modeling of Topic Dynamics with Multiple Timescales [J]. ACM Transactions on Knowledge Discovery from Data, New York: ACM Press, 2012, 5(4): 19.

[17] Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends [C]// Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM Press, 2006: 424-433.

[18] Minka T, Lafferty J. Expectation-propagation for the generative aspect model[C]// Proceedings of the 18th UAI conference on Uncertainty in artificial intelligence, San Francisco, Morgan Kaufmann Publishers, 2002: 352-359.

[19] Minka T. Expectation propagation for approximate Bayesian inference[C]// Proceedings of the 17th UAI conference on Uncertainty in artificial intelligence, San Francisco, Morgan Kaufmann Publishers, 2001: 362-369.

[20] Griffths T, Steyvers M. Finding scientific topics[C]// Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Suppl. 1): 5228-5235.

[21] Teh Y, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation[J]. Advances in neural information processing systems, 2007, 19: 1353-1360.

[22] Bishop C. Pattern recognition and machine learning[M]. New York: Springer-Verlag, 2006.

[23] Owen C B. Parameter Estimation for the Beta Distribution [D]. Provo: Brigham Young University, 2008.

[24] Azzopardi L, Girolami M, Van Risjbergen K, et al. Investigating the relationship between language model perplexity and IR precision-recall measures [C]// Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, New York: ACM Press, 2003: 369-370.

(责任编辑　马　兰)