北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.035

一种结合有监督学习的动态主题模型

蒋卓人[1]　陈燕[1,†]　高良才[2]　汤帜[2]　LIU Xiaozhong[3]

1. 大连海事大学交通运输管理学院，大连　116026; 2. 北京大学计算机科学技术研究所，北京　100871;
3. Department of Information and Library Science, Indiana University Bloomington, Bloomington, IN, USA 47405;
†通信作者, E-mail: chenyan_dlmu@163.com

摘要　针对传统主题模型存在的不足，提出一种新的结合有监督学习的动态主题模型(Supervised Dynamic Topic Model, S-DTM)。该模型不仅能够随时间的变化对语言进行动态建模，而且结合有监督学习技术，在主题变分推理中加入了标签约束，从而建立主题与标签之间的映射关系，提高了主题的表达解释能力。通过在一个跨越 25 年的“以自然语言处理领域的中文期刊论文为主导”的中文语料库上的实验，证明该有监督的动态主题模型相较于静态的有监督主题模型和无监督的动态主题模型，具有更好地语义解释概括能力，更能准确反映文档的主题结构，更能精确捕捉主题–词汇概率分布的动态演化。

关键词　有监督学习；动态主题模型；变分推理
中图分类号　TP18

A Supervised Dynamic Topic Model

JIANG Zhuoren[1], CHEN Yan[1,†], GAO Liangcai[2], TANG Zhi[2], LIU Xiaozhong[3]

1. College of Transportation Management, Dalian Maritime University, Dalian 116026; 2. Institute of Computer Science & Technology, Peking University, Beijing 100871; 3. Department of Information and Library Science, Indiana University Bloomington, Bloomington, IN, USA 47405; †Corresponding author, E-mail: chenyan_dlmu@163.com

Abstract　An innovative Supervised Dynamic Topic Model (S-DTM) is developed for overcoming the limitation of tranditional topic models. S-DTM models the time-varying language dynamics and is combined with supervised learning technology by adding label restriction in topic variational inference. It makes the topic-label mapping and improves the interpret ability of topics. A set of experiments is conducted on a twenty-five-year-spanning Chinese journal paper corpus that is mainly focusing on natural language processing. Experiment results show that compared with static supervised topic model and unsupervised dynamic topic model, S-DTM has a better semantic interpretation performance, reflects the topic structure of a document more accurately, captures the dynamic evolution of the term-distribution of topics more precisely.

Key words　supervised learning; dynamic topic model; variational inference

随着信息技术不断发展，大量的信息资源不断涌现，新闻、博客、网页、科技文献、书籍、社交网络信息等包围了人们的生活。面对海量信息，传统的信息获取技术已经无法满足人们的信息需求，亟需新的技术帮助人们自动地组织、索引、搜索和浏览这些海量数据集。概率主题模型(Probabilistic Topic Models)是近年来在机器学习和统计学领域中广泛研究和应用的新技术，这是一系列旨在发现和标记大规模文档的主题信息的算法[1]。概率主题模型对于海量语料库是一种有效的降维方式，在自然语言处理、信息检索、文本挖掘等诸多方面已经有了较为成功的应用[2]。

然而，目前大部分的概率主题模型都基于假设–语料库中的文档之间是没有顺序的，换言之，语

国家自然科学基金 (61202232, 71271034)和中央高校基本科研业务费(3132014307)资助
收稿日期: 2014–06–30; 修回日期: 2014–10–22; 网络出版时间: 2014-12-01 09:26

料库中的文档是可交换的。这种简化的假设实际上不够恰当且不符合实际情况: 1) 大量的语料库比如科技文献库、新闻文本库都是时序语料库, 它们中的文档都是有时间属性的, 一些特定文本信息只可能出现在某个特定的时间段内。另外, 不少语料库跨越了上百年的时间, 对于这些语料库, 忽视时间顺序属性显然不恰当; 2) 语言是随着时间变化的, 作为对文档文本的"主题性的总结", 主题本身必然也将随着时间不断演化。如图 1 所示, 本文比较了不同时期两位美国总统的就职演说, 并以词汇云(Word Cloud)的方式加以展现, 出现频率越高的词汇, 其展现尺寸越大。从图中可以明显看出, 尽管都是美国总统的就职演说, 但是从 1789 年到 2013 年, 他们使用的词汇不尽相同, 并且高频词汇的运用也是大相径庭。因此, 当语料库的时间跨度较大的时候, 主题模型显然需要考虑时间因素和语言的动态变化。所以, 如何在主题模型中加入语言随着时间变化的特性, 从而更为准确地对主题动态演化进行建模, 是一个值得研究的问题。

另一方面, 在主题模型中, 主题是一个语义层面的潜在变量, 一个主题就是一系列词汇的概率分布(term-distribution), 目前大量主流的主题模型都是无监督的, 这就导致: 1) 学习得到的主题往往解释性较差, 有时候很难被理解; 2) 由于主题模型大都假设主题数目已知并且固定, 那么, 在主题是潜在的情况下, 如何准确而快速地为不同的语料库确定一个恰当的主题数目也成为一个棘手的问题。

针对上述问题, 本文提出一种结合有监督学习的动态主题模型(Supervised Dynamic Topic Model, S-DTM), 该模型是应用于多标签的时序语料库上的一个时间序列主题模型(Time-series topic model), 它基于动态主题模型(Dynamic Topic Model, DTM)[3], 能够允许不同主题下的词汇概率分布随着时间演化, 并且结合附加标签的隐含狄利克雷分配 (Labeled-LDA, L-LDA)[4], 在模型中加入有监督学习技术, 利用文档自身的元数据(meta-data)——标签(label)进行监督学习, 建立主题与相应标签的映射关系, 用标签表示主题, 解决了主题难以表达解释和主题数目难以确定的问题。我们在一个跨越 25 年(1985—2009)的以自然语言处理领域的中文期刊论文为主导的 DBLP 中文语料库[5]上, 与静态的有监督主题模型和无监督的动态主题模型进行对比实验, 证明 S-DTM 在语义解释概括、主题结构描述和动态语言变化的捕捉等方面具有更好的性能。

1 相关工作

作为目前自然语言处理和信息检索领域的重要课题和研究热点, 概率主题模型是一系列有效发掘隐藏在大规模文档中的主题结构的算法, 这种主题结构为探索和理解文档打开了一扇新的窗口。作为一种统计方法, 它通过分析文本中的词来发现蕴藏于文档中的语义主题结构, 从而实现对文本的组织和归纳[1]。在主题模型中, 主题是一个语义层面的假设, 一个主题是一系列词汇的概率分布, 而文档可以表示为一系列主题的混合。隐含狄利克雷分配(Latent Dirichlet Allocation, LDA)[6]是最简单和经典的主题模型。通过训练, 能够得到每一个隐含主题上的词汇概率分布, 而通过在文档层面上的后验推

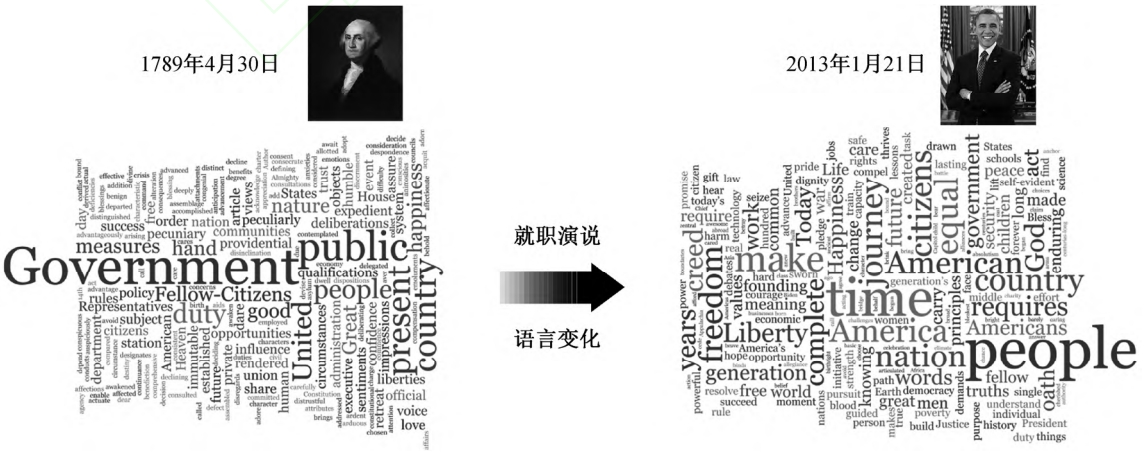

图 1 美国总统就职演说词汇云

Fig. 1 Word cloud of inaugural addresses of the USA Presidents

理(posterior inference), 又能获得不同主题在文档上的分布比例。目前 LDA 已经在学术界和工业界得到广泛研究和应用, 并且大量基于 LDA 的更为复杂的主题模型也在近几年的研究中不断提出[7–9]。尽管在研究和应用中取得广泛成功, LDA 及其为基础的主题模型也存在一些对现实情况过于简化的假设, 如何弱化和扩展这些假设以发现文本中更加复杂的结构, 成为当前主题建模研究领域中的一个研究热点[1]。

正如前文所论述, 大多数研究都忽视了语料库中文本的时间顺序, 在语料库时间跨度较大时, 这种简化假设变得极为不适应。比如在科技论文的语料库中, 一篇发表于 1910 年的主题为“理论物理”的论文和一篇发表于 2010 年的相同主题的论文, 它们的主题词汇分布有着巨大差别。这就需要主题模型算法不仅能够发掘文本的主题结构, 同时也能够捕捉到主题结构随着时间的动态演变。Blei 等[3]提出动态主题模型 (Dynamic Topic Model, DTM), 考虑了文档的先后顺序, 假设主题随着时间而发生变化, 并给出比 LDA 内涵更丰富的后验主题结构, 该模型也被证明是一个能够准确描述潜在主题及其动态变化的强有力的工具。文档影响力模型[10](Document Influence Model, DIM)是一种基于 DTM 的主题模型, 在 DIM 中, 每篇文档分配一个正态分布的影响力分数, 从文本角度该影响力分数显示该文档对潜在主题的影响力。该模型也已被证明是一个实用的评估文档主题动态影响力的模型。

尽管在主题动态演化方面, DTM 取得了成功, 但是在实际应用中还是存在着诸多不便, 作为无监督模型, DTM 也存在着和其他无监督主题模型类似的问题。其中一个重要的问题就是主题的解释问题, 通过无监督训练得到的主题有时非常难以理解, 在语义理解角度失去了解释的功能。另一个问题就是如何确定语料库中的主题数目, 因为一个恰当的主题数目控制着主题的“粒度”, 对最终主题建模的准确性和有效性有重大影响。而且大多数主题模型均假设主题的数量已知且固定, 这就意味在实验开始前就需要设定语料库的主题数目, 使得如何准确地确定主题的数目成为一个棘手的问题。当前也有很多研究工作围绕这两个问题展开, 文献[13]提出一种附加标签的主题模型, 它将一个类别标签变量分配给一篇文档, 而这个标签变量本身是一个主题分布; 文献[11]提出一种贝叶斯非参数模型, 在该模型中, 主题数目是在对文档集的后验推理中确定的, 新的文档可以扩展之前没有的主题。通过数据推理, 该数据模型将主题结构扩展成层次结构。在实际操作中, 常通过比较一组不同的主题数目, 选取获得最大似然概率的主题数目作为最终主题数目。以上这些方法都不能简单有效地确定主题数目或者解释主题。另一方面, 现实中大量的语料库存在标签元数据, 比如科技论文语料库中的关键词等, 这些标签往往由作者提供, 能较好地概括文档所涵盖的内容与主题, 具有很强的解释性。附加标签的隐含狄利克雷分配[4]是由斯坦福大学的自然语言处理小组提出的一种扩展了 LDA 的有监督主题模型, 它将主题与标签联系起来, 不同于 Supervised LDA[12]和 DiscLDA[13]假设文档只可以与单一的标签相关联, L-LDA 可以成功应用于多标签语料库, 因此每个文档可以拥有多个标签。L-LDA 引入有监督的学习技术, 将文档的元数据标签作为主题标签, 从而能将文档中的词汇与该文档所拥有的主题标签联系起来。L-LDA 成功地解决了主题的表示问题, 并且由于元数据的标签数目是固定的, 因此可以方便地确定主题数目。

本文正是在 DTM 的基础上结合 L-LDA, 通过引入有监督的学习技术, 将主题与数据集本身的元数据标签相映射, 从而把无监督的主题模型扩展为有监督的动态主题模型。

2 结合有监督学习的动态主题模型

2.1 模型概述

本文开发了一种能够应用于多标签时序语料库的结合有监督学习的动态主题模型, 在 DTM 的基础上结合 L-LDA, 引入有监督的学习技术, 将 DTM 扩展为 S-DTM, 使模型不仅能够准确捕捉语言在不同时间片上的变化, 也能够将潜在的主题利用文档自身的元数据标签而显式表现, 对主题的“粒度”实现了更准确地控制, 解决了主题数目难以确定的问题, 也使得主题结构的可解释性大大增强。S-DTM 图形化表示见图 2。

本文规定语料库 D 中的文档共含有 V 个可唯一识别的词汇(term), 按照文档的时间属性将语料库分为 T 个时间片, 时间片 1 是最久远的时间片, 时间片 T 是最新近的时间片, 每个时间片上的文档子集为$\{D_1, \ldots, D_T\}$。

语料库共存在 K 个唯一识别的元数据标签

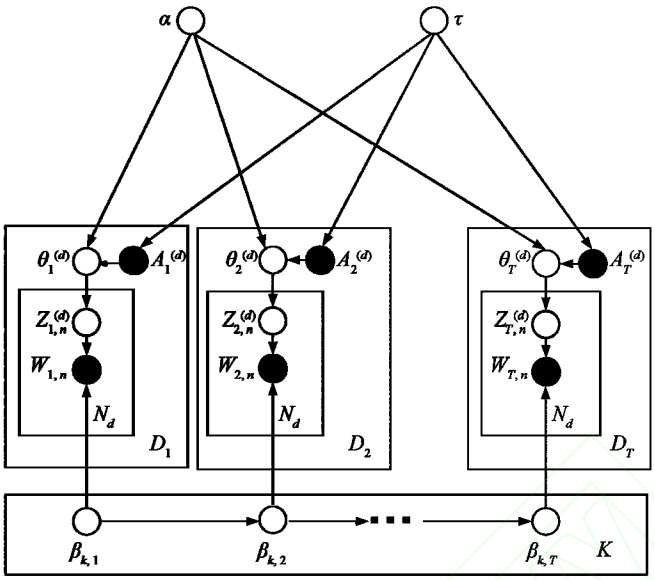

图 2　结合有监督学习的动态主题模型的图形化表示

Fig. 2　Graphical representation of a supervised dynamic topic model

(label), 将 K 个标签分别与 K 个主题相映射, 每个主题的涵义即是其映射的标签, 从而解决了主题的解释问题。语料库中的一篇文档 d, 可以表示为一系列词(word)的元组 $w^{(d)}=(w_1,\ldots,w_{N_d})$ 和一系列的二元的主题"存在/不存在"的标识变量 $\Lambda^{(d)}=(l_1,\ldots,l_K),l_k\in\{0,1\}$。$l_k=0$ 表示 d 不存在第 k 个标签, 从而 d 与第 k 个主题无关, 类似地, $l_k=1$ 表示 d 存在该标签, 从而 d 与第 k 个主题相关。例如, 将关键词元数据作为标签, 如果语料库中共有关键词 5 个, 我们就把一个关键词对应一个主题(主题–标签映射), 这样就确定训练集共有 5 个主题, 假设文献 A 拥有关键词 1 和关键词 3, 那么 Λ(A) = (1, 0, 1, 0, 0), 即文献 d 仅仅与主题 1 和主题 3 相关。对于 d 中的每一个词 w_n, 也都可以根据 $\Lambda^{(d)}$ 来确定其是否与第 k 个主题相关, N_d 表示文档 d 的长度。

在主题模型中, 每个主题是一个词汇概率分布(term-distribution), 用β 来表示。但是与之前大部分的研究不同, S-DTM 并不存在固定的主题–词汇分布, 而是引入一个多元对数正态马尔可夫链, 来捕捉不同时间片上的语言动态变化, 自然参数(natural parameter) $\beta_{k,t}$是一个 V 维向量, 表示在时间片 t 上的第 k 个主题, 其中, $\beta_{k,t}$的第 w_n维分量可以由均值参数π表示为 $\beta_{k,t,w_n}=\log(\pi_{w_n}/\pi_V)$[3]。主题–词汇概率分布状态随着时间演变可以表示为

$$\beta_{k,t}\mid\beta_{k,t-1}\sim\mathbb{N}(\beta_{k,t-1},\sigma^2I)。\tag{1}$$

需要注意的是, 尽管 S-DTM 将主题与语料库中的元数据标签相关联, 但在不同的时间片上, 即便主题与同一个标签相关联, 但它们的词汇分布依旧不一样, 是随着时间而演化的。例如, 在时间片"1990 年至 1992 年"上的主题"机器学习"和时间片"2005 年至 2006 年"上的主题"机器学习", 它们的词汇分布是不同的。

对于每个时间片上的文档子集, 本文把它们用 L-LDA 建模。在 S-DTM 中, 时间片 t 上的文章生成过程如下。

1 循环该时间片上的每一个主题 $k\in\{1,\ldots,K\}$

2 根据 $\beta_{k,t}\mid\beta_{k,t-1}\sim\mathbb{N}(\beta_{k,t-1},\sigma^2I)$ 生成该时间片上的主题–词汇概率分布参数变量 $\beta_{k,t}$

3 循环每一篇位于该时间片上的文档 d

4 循环该时间片上的每一个主题 $k\in\{1,\ldots,K\}$

5 根据 $\Lambda_{t,k}^{(d)}\in\{0,1\}\sim\text{Bernoulli}(\cdot\mid\tau_{t,k}^{(d)})$, 生成二元主题标识变量 $\Lambda_{t,k}^{(d)}$

6 根据 $\alpha_t^{(d)}=L_t^{(d)}\times\alpha_t$, 生成 Dirichlet 主题先验分布的参数变量 $\alpha_t^{(d)}$

7 根据 $\theta_t^{(d)}=(\theta_{m_1^{(d)}},\ldots,\theta_{m_{M_d}^{(d)}})^T\sim Dir(\cdot\mid\alpha_t^{(d)})$, 生成多项分布参数变量 $\theta_t^{(d)}$

8 循环文档 d 中的每一个词 $w_{t,n},n\in\{1,\ldots,N_d\}$:

9 根据 $z_{t,n}^{(d)}\in\{\lambda_{m_1^{(d)}},\ldots,\lambda_{m_{M_d}^{(d)}}\}\sim\text{Mult}(\cdot\mid\theta_t^{(d)})$, 生成词–主

题分配标识向量 $z_{t,n}^{(d)}$

10 根据 $w_{t,n} \in \{1,\ldots,V\} \sim \text{Mult}(\cdot \mid \pi(\beta_{z_{t,n}^{(d)},t}))$，生成词 $w_{t,n}$

在如上所示的生成过程中，每一行代表一个生成步骤。标号 1 表示步骤 1，以此类推。步骤 1 和 2 表明，每个时间片 t 上的主题 $\beta_{k,t}$ 是由前一个时间片 t–1 上对应的主题 $\beta_{k,t-1}$ 平滑演化而来，这一步与传统的 DTM 相同。接下去的步骤，传统无监督的 DTM 将根据狄利克雷(Dirichlet)先验分布参数 α_t 生成一个文档上的主题分布，即 K 个主题的多项分布参数变量 θ_t。然而，与 DTM 不同，S-DTM 在下面的步骤结合 L-LDA，通过二元主题标识向量 $\Lambda_t^{(d)}$，将 $\theta_t^{(d)}$ 限制为仅含有该文档拥有的元数据标签对应的主题。

步骤 5 为 S-DTM 利用伯努利分布生成处于时间片 t 上的每个文档的二元主题标识向量 $\Lambda_t^{(d)} = (l_1,\ldots,l_K), l_k \in \{0,1\}$。$\tau_{t,k}^{(d)}$ 表示文档标签的先验概率，根据数据集自身的标签元数据就可以获得。

步骤 6 为 S-DTM 计算一个文档相关的标签约束矩阵 $L_t^{(d)}$，对于每篇位于时间片 t 上的文档来说，$L_t^{(d)}$ 是一个 $M_d \times K$ 的矩阵，根据 $\Lambda_t^{(d)}$ 中分量值为 1 的分量生成一个新的标签向量：$M_t^{(d)} = \{m_i^{(d)} = k \mid l_k = 1\}, l_k \in \Lambda_t^{(d)}$，$M_d = \left|M_t^{(d)}\right|$。$L_t^{(d)}$ 的计算方法如下:

$$L_{i,j}^{(d)} = \begin{cases} 1, & m_i^{(d)} = \text{j}, \\ 0, & \text{其他情况}, \end{cases} \tag{2}$$

其中，$0 \leqslant i \leqslant M_d, 0 \leqslant j \leqslant K$。利用 $L_t^{(d)}$ 将 K 维狄利克雷(Dirichlet)主题先验分布参数 α_t 降维到 M_d 维的 $\alpha_t^{(d)}$ 分布参数:

$$\alpha_t^{(d)} = L_t^{(d)} \times \alpha_t = (\alpha_{m_1^{(d)}},\ldots,\alpha_{m_{M_d}^{(d)}})^T。 \tag{3}$$

假设语料库中主题标签数 K=3，时间片 t 上的文档 d 拥有第 1 个标签和第 3 个标签，因此 $\Lambda_t^{(d)} = (1,0,1)$，从而 $M_t^{(d)} = \{1,3\}, M_d = 2$，$L_t^{(d)}$ 即为 $\begin{pmatrix} 1,0,0 \\ 0,0,1 \end{pmatrix}$。又 $\alpha_t = \{\alpha_1,\alpha_2,\alpha_3\}^T$，则 $\alpha_t^{(d)} = L_t^{(d)} \times \alpha_t = \{\alpha_1,\alpha_3\}^T$，降维完成。降维的狄利克雷(Dirichlet)主题先验分布参数 $\alpha_t^{(d)}$ 被约束为只与标签 1 和标签 3 相关。

步骤 7 是根据已被标签约束的降维狄利克雷(Dirichlet)主题先验分布参数 $\alpha_t^{(d)}$，生成文档的主题多项分布参数 $\theta_t^{(d)}, \theta_t^{(d)}$ 是文档的主题分布参数，表示不同主题在文档中所占的比例，$\theta_t^{(d)}$ 受标签约束。

步骤 9 中 $z_{t,n}$ 表示第 n 个词 $w_{t,n}$ 的主题分配标识向量，t 表示时间片，当第 k 个分量 λ_k 值为 1 时，表示 $w_{t,n}$ 是根据第 k 个主题–词汇概率分布 $\beta_{k,t}$ 生成。由于 $z_{t,n}^{(d)}$ 是根据 $\theta_t^{(d)}$ 分布生成的，因此当 $\theta_t^{(d)}$ 增加标签约束后，$z_{t,n}^{(d)}$ 也只会根据文档自身所拥有的 M_d 个标签进行分配。这样，S-DTM 就可以保证文档 d 中的所有词都只由该文档所拥有的标签所对应的主题生成。

步骤 10 中 $\beta_{z_{t,n}^{(d)},t}$ 表示基于上一步骤中生成的 $z_{t,n}$ 为第 n 个词 $w_{t,n}$ 分配的主题，即该主题的词汇概率分布，根据 $\beta_{z_{t,n}^{(d)},t}$ 即可生成文档中的词汇。$\beta_{z_{t,n}^{(d)},t}$ 由均值参数 π[3] 表示：$\pi(\beta_{z_{t,n}^{(d)},t})_{w_{t,n}} = \dfrac{\exp(\beta_{z_{t,n}^{(d)},t,w_{t,n}})}{\sum \exp(\beta_{z_{t,n}^{(d)},t,w_{t,n}})}$。

2.2 变分推理

在 S-DTM 中，每个位于时间片 t 上的文档都被建模为 L-LDA 模型，而文档–词汇概率分布则随着时间动态变化，建模为一条多元对数正态马尔可夫链。在文档层面上，S-DTM 包含的隐含变量是文档的主题部分参数 $\theta_t^{(d)}$ 和文档中词–主题分配标识向量 $z_t^{(d)}$；在主题层面上，S-DTM 包含的隐含变量是时间片 t 上的主题–词汇分布参数变量 $\beta_{k,t}$。

本文的目标是在多标签时序语料库的训练样本基础上，对包含隐含参数变量的后验分布进行推理(inference)。高斯模型(Gaussian models)能够被应用于时间序列上的自然参数的处理，从而对于时间上的动态变化进行建模，然而，由于高斯分布和多元分布的非共轭性(Nonconjugacy)，直接对后验分布进行推理难以操作。另一方面，尽管吉布斯抽样(Gibbs sampling)在静态的主体模型上得到有效的应用[21]，但是非共轭性导致抽样方法(Sampling methods)很难应用于动态模型[3]。因此本文采用变分推理[14] (Variational Inference)来近似后验分布。变分推理首先假设一个更为简单的包含隐含变量的分布，这个分布包含相应的变分变量，通过不断更新变分变量，达到不断优化变分分布与真实后验分布之间的 KL 分歧(Kullback-Liebler Divergence)的目的，当 KL 分歧小于某个阈值后，即可将变分分布作为真实的后验分布的近似替代。S-DTM 中:

1) 属于文档层面的隐含变量有两个，$\theta_t^{(d)}$ 被赋予一个的带有 $\gamma_t^{(d)}$ 参数变量的 Dirichlet 分布，而 $z_t^{(d)}$ 被赋予一个带有 $\phi_t^{(d)}$ 参数变量的多项式分布。对于每一篇存在于特定的时间片 t 上的文档 d 来说，实际的后验形式为

$$p(\theta_t^{(d)}, z_t^{(d)} \mid w^{(d)}, \alpha_t^{(d)}, \beta_t) = \frac{p(\theta_t^{(d)}, z_t^{(d)}, w^{(d)} \mid \alpha_t^{(d)}, \beta_t)}{p(w^{(d)} \mid \alpha_t^{(d)}, \beta_t)},$$

本文构建的文档层面的变分分布如下:

$$q(\theta_t^{(d)}, z_t^{(d)} \mid \gamma_t^{(d)}, \phi_t^{(d)}) = q(\theta_t^{(d)} \mid \gamma_t^{(d)}) \prod_{n=1}^{N_d} q(z_{t,n}^{(d)} \mid \phi_{t,n}^{(d)}) 。 \quad (4)$$

2) 在主题层面，S-DTM 对隐含变量 β_t 的建模如下:

$$\begin{aligned} \beta_t \mid \beta_{t-1} &\sim \mathbb{N}(\beta_{t-1}, \sigma^2 I), \\ w_{t,n} \mid \beta_t &\sim \text{Mult}(\pi(\beta_t)), \\ \hat{\beta}_t \mid \beta_t &\sim \mathbb{N}(\beta_t, \hat{v}_t^2 I)。 \end{aligned}$$

β_t 对应的变分变量是 $\hat{\beta}_t$，它们被解释为正态分布链上的高斯变分观测变量[3](Gaussian "variational observations")，$\hat{v}_t$ 是变分方差变量，本文利用变分分布 $q(\beta_{1:T} \mid \widehat{\beta_{1:T}})$ 来近似实际后验分布 $p(\beta_{1:T}|D)$，其变分分布的形式为

$$q(\beta \mid \hat{\beta}) = \prod_{k=1}^{K} q(\beta_{k,1:T} \mid \hat{\beta}_{k,1:T}), \quad (5)$$

其中，D 是时序语料库训练样本，$\beta_{k,1:T}$ 是 $\{\beta_{k,1}, \ldots, \beta_{k,T}\}$，表示不同时间片上第 k 个主题变量，$\widehat{\beta_{k,1:T}}$ 是 $\widehat{\beta_{k,1}}, \ldots, \widehat{\beta_{k,T}}$，表示不同时间片上第 k 个主题变量对应的高斯变分观测变量。

最终，本文确定的模型变分分布为

$$q(\beta, \theta, z \mid \hat{\beta}, \gamma, \phi) = \prod_{k=1}^{K} q(\beta_k \mid \widehat{\beta_k}) \prod_{t=1}^{T} \left(\prod_{d=1}^{D_t} (q(\theta_t^{(d)} \mid \gamma_t^{(d)}) \prod_{n=1}^{N_d} q(z_{t,n}^{(d)} \mid \phi_{t,n}^{(d)})) \right), \quad (6)$$

其中，D_t是时间片 t 上的文档子集。在变分推理中，最小化变分分布与真实分布之间的 KL 分歧等同于最大化样本的边缘概率[16]，在 S-DTM 中，目标就是要最大化下式的下界:

$$\begin{aligned} \ln p(D) \geqslant & \sum_T \mathbb{E}_q[\ln p(\beta_t \mid \beta_{t-1})] + \\ & \sum_T \sum_{D_t} \mathbb{E}_q[\ln p(\theta_t^{(d)} \mid \alpha_t^{(d)})] \\ & + \sum_T \sum_{D_t} \sum_{N_d} \mathbb{E}_q[\ln p(z_{t,n}^{(d)} \mid \theta_t^{(d)})] + \\ & \mathbb{E}_q[\ln p(w_{t,n} \mid z_{t,n}^{(d)}, \beta_t)] + H(q)。 \end{aligned} \quad (7)$$

该下界通过坐标上升法(coordinate ascent)优化，变分变量按照不同时间片被分别优化。式(7)的下界的相对变化小于设定的阈值后，变量的迭代更新才会停止。

在 S-DTM 的变分推理中，对隐含变量 $\theta_t^{(d)}$ 和 $z_t^{(d)}$ 的推理转化为对 $\gamma_t^{(d)}, \phi_t^{(d)}$ 的迭代更新，其迭代更新的方程为

$$\begin{aligned} \phi_{t,n,i}^{(d)} &\propto \beta_{t,n,i} \exp(\mathbb{E}_q[\log \theta_{t,i}^{(d)} \mid \gamma_t^{(d)}]) \\ \gamma_{t,i}^{(d)} &= \alpha_{t,i}^{(d)} + \sum_{n=1}^{N_d} \phi_{t,n,i}^{(d)}, \end{aligned} \quad (8)$$

其中，$\phi_t^{(d)}$ 的更新中的 $\mathbb{E}_q[\log \theta_{t,i}^{(d)} \mid \gamma_t^{(d)}]$ 根据如下的方法计算:

$$\mathbb{E}_q[\log \theta_{t,i}^{(d)} \mid \gamma_t^{(d)}] = \Psi(\gamma_{t,i}^{(d)}) - \Psi(\sum_{j=1}^{K} \gamma_{t,j}^{(d)}),$$

其中，Ψ是函数 $\log\Gamma$ 的一阶导函数。

从形式上看，S-DTM 在文档层面上的变分推理迭代更新过程与传统的 LDA 非常相似，但是，两者是明显不同的，由于 S-DTM 是可监督的，所以更新过程中加入了可监督的学习技术。每个时间片 t 上文档的二元主题标识变量 $\Lambda_t^{(d)}$ 都可以通过观察得到，因此，与普通的无监督主题模型不同，S-DTM 的主题先验分布参数 $\alpha_t^{(d)}$ 是一个受到标签约束的变量，这使得在每篇文档上: 1) 迭代更新的过程与文档自身拥有的主题标签有关; 2) 只有与文档标签相关的 $\gamma_t^{(d)}$ 和 $\phi_t^{(d)}$ 才能得到迭代更新，从而模型将目标文档的主题份额控制为只向文档所拥有的主题标签上进行分配。

在主题层面上，S-DTM 对隐含变量 $\beta_{1:T}$ 的变分推理与传统 DTM 相同，其对变分变量 $\widehat{\beta_{1:T}}$ 的更新，采用基于卡尔曼过滤算法[15](Kalman filter)的近似推理[3]。

3 实验与结果分析

3.1 实验数据

本文采用一个时间跨度为 25 年(1985—2009 年)以自然语言处理领域中文期刊论文(DBLP 资源)为主导的中文语料库[5]进行实验。该语料库借助自动化学科知识服务平台的知识族谱进行构建，以“自然语言处理”、“人工智能”、“数据挖掘”和“信息检索”等专业术语为查询节点，利用有限步拓展从自动化学科知识服务平台数据中抽取资源，并将抽取的子资源构建为语料库。

该语料库共包含论文 4615 篇，选取论文的中文摘要作为文档的文本描述信息。本文将论文的关

键词作为标签, 其优势主要有: 1) 关键词是由大量不同的作者提供, 对文章所涵盖的主题的概括更为准确和全面, 由此确定的主题数目相较于传统主题模型主观确定的数目也更为自然、合理; 2) 由关键词作为标签, 主题在语义表现层面上也更有解释性。本文首先通过人工分析, 将意义相近的关键词合并作为同一个关键词, 随后选取了 65 个关键词作为主题标签, 所有被选中的关键词至少出现在 20 篇论文内。语料库的词汇数目为 7556, 所有被选中的词汇必须满足以下两个条件: 1) 不是最常出现的前 5%的词汇; 2) 每个词汇至少出现在两篇文档中。本文将语料库分为 10 个时间片, 每个时间片上的论文数大致相当, 如表 1 所示。

3.2 实验方案

为了评价 S-DTM 的在语义方面的解释概括性能, 本文将 S-DTM 应用于文本分类任务, 选取静态的 L-LDA 和传统的无监督 DTM 作为对照实验组。

1) 在训练集上, 分别使用 S-DTM, L-LDA 和 DTM 进行训练, 通过基于关键词标签约束的监督训练学习, S-DTM 能够得到一组不同时间片上的主题-词汇概率分布 $\beta_{1:T}^{S-\mathrm{DTM}}$, 每一个时间片 t 上的 $\beta_t^{S-\mathrm{DTM}}$ 都是一个 $K\times V$ 的矩阵, K=65, 是关键词的个数, V=7556, 是词汇表的数目。通过无监督训练学习, DTM 得到一组不同时间片上的主题–词汇概率分布 $\beta_{1:T}^{\mathrm{DTM}}$, 本文将主题数目预先设定为 65 个, 以保证 β_t^{DTM} 也是一个65×7556的矩阵。通过监督训练学习, L-LDA 得到的是一个固定不变的全局主题–词汇概率分布 $\beta^{L-\mathrm{LDA}}$, 类似的, $\beta^{L-\mathrm{LDA}}$ 也是一个 65×7556 的矩阵。

2) 根据训练得到的主题–词汇概率分布, 在文档层面上统一采用经典的 LDA 后验推理[6]对测试集内的论文进行推理, 将测试文档转化为一个 65 维的主题向量。其中, 对于 S-DTM, 将根据测试论文的发表时间, 基于相应时间片上的 β_t^{S-DTM} 进行推理; 类似地, 对于 DTM, 也将基于相应时间片上的 β_t^{DTM} 进行推理; 由于 L-LDA 的主题–词汇概率分布是固定的, 因此推理总是基于同一个 $\beta^{L-\mathrm{LDA}}$ 进行。

3) 将推理生成的主题向量作为该文档的特征值向量, 应用于分类算法中来预测该文档的类别。

为了防止文档主题特征向量生成结果产生“偏歧”, 在针对以上 3 种模型的训练和测试中, 本文都采用了 5 折交叉验证(5-fold cross-validation)。

在最终推理得到的 4615 文档中, 本文选择 2072 个文档, 根据其自身的关键词标签, 把文档集分成 13 类, 进行文档分类预测。其中, 每一个被选中的文档有且仅有一个由作者标注的关键词, 而被选中的关键词标签至少出现在 60 篇以上的文档中。

本实验中, 3 种经典的机器学习算法被应用于文本分类: 1) 朴素贝叶斯(Naïve Bayes, NB); 2) 最大熵(Maximum Entropy, ME); 3) C4.5 分类决策树(C4.5)。为了防止文本分类的预测结果出现“偏歧”, 这 3 种机器学习算法在训练和测试中, 采用 10 折交叉验证(10-fold cross-validation)。

文本采用的分类评价指标主要有两项: 准确率(Accuracy)和 F1 值(F1-measure):

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}},$$

$$\mathrm{F1}=2\times\frac{\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}},$$

其中, TP 表示正确判断为该类的数目(true positive), TN 表示正确判断不为该类的数目(true negative), FP 表示错误判断为该类的数目(false positive), FN 表示错误判断不为该类的数目(false negative), precision 表示精确率, recall 表示召回率。

本实验的主要目的并不是追求文本分类的性能表现, 而是通过比较不同主题模型生成的主题特征向量对文本分类性能产生的影响, 从而比较它们对文本的语义解释概括能力。

3.3 实验工具

实验使用常用的文本挖掘工具和开源代码: 1) 中文分词方面, 采用 mmseg4j 中文分词器[17], 该分词器实现了 MMSeg 算法; 2)文本分类算法方面,

表 1 时间片划分

Table 1 Time slice partition

时间片	文档数	时间片	文档数
1996 年前	392 篇	2005 年	426 篇
1996–1999 年	402 篇	2006 年	511 篇
2000–2001 年	413 篇	2007 年	506 篇
2002–2003 年	548 篇	2008 年	531 篇
2004 年	438 篇	2009 年	448 篇

采用了 Mallet toolkit[18]; 3) DTM 的代码采用的是 DTM 作者分享的源代码[19], S-DTM 的开发也基于此代码; 4) L-LDA 的代码采用的是开源项目 JGibbLabeledLDA[20]。

3.4 实验结果和分析

将 L-LDA, DTM 和 S-DTM 生成的主题特征向量应用于不同的分类机器学习算法, 最终结果如图 3 所示, 具体数据见表 2。

表 2 不同主题特征向量的文本分类性能数据

Table 2 Text categorization performance on different topic feature vectors

机器学习算法	主题模型	准确率	F1 值
朴素贝叶斯 Naïve Bayes	L-LDA	0.3084	0.1397
	DTM	0.3113	0.1137
	S-DTM	0.3557*	0.1876*
最大熵 Maximum Entropy	L-LDA	0.5333	0.4558
	DTM	0.4749	0.3836
	S-DTM	0.5497*	0.4929*
C4.5 决策树(C4.5)	L-LDA	0.4817	0.4670
	DTM	0.4223	0.3913
	S-DTM	0.5164*	0.4780*

说明: * 为最优结果。

通过图 3 和表 2 得到以下结论。

1) 不论是应用何种机器学习算法进行文本分类, 由 S-DTM 生成的主题特征向量在准确度和 F1 值这两项指标上总能取得最优的分类效果。这表明, 相较于静态的 L-LDA 和无监督的 DTM, 结合有监督学习的动态主题模型 S-DTM 对于文档的语义概括解释能力最强, 由 S-DTM 生成的主题向量最能反映文档的真实主题结构。

2) 在绝大部分指标上, DTM 的表现都是最差的, 这是因为 L-LDA 和 S-DTM 都是有监督的主题模型, 由于在模型训练学习中加入了主题标签约束, 因此, 对于文档的主题结构的学习更为准确。而 DTM 在无监督的情况下, 尽管主题数目一致, 但是主题的“粒度”是不可控的, 主题本身涵盖的范围也是未知的, 由此导致 DTM 生成的主题结构可能不是最为合理的。

3) 尽管在训练学习过程中对标签进行了约束, 但是 S-DTM 依旧在各项指标上超过了 L-LDA, 这是因为 S-DTM 是动态的主题模型, 其主题–词汇分布随着时间而不断改变, 而 L-LDA 是一个静态主题模型, 只有一个全局的固定的主题–词汇分布, 因

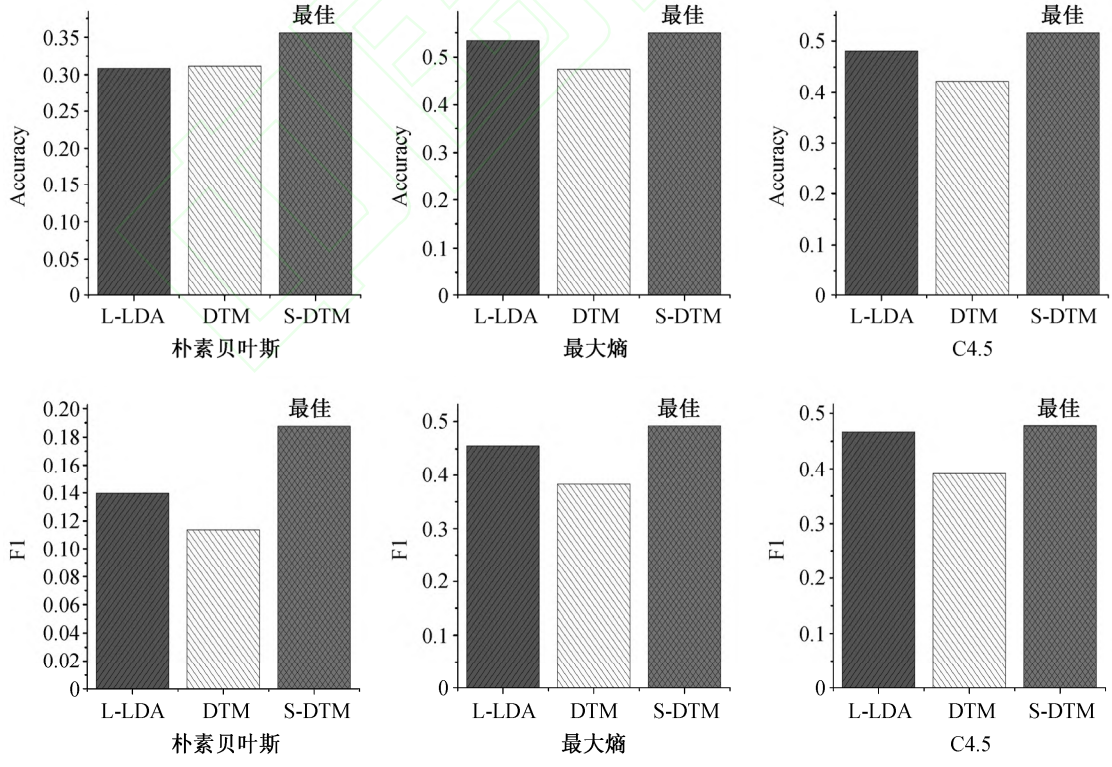

图 3 不同主题特征向量的文本分类性能表现

Fig. 3 Text categorization performance on different topic feature vectors

此, S-DTM 对主题–词汇的概率分布的建模更准确。

图 4 和 5 进一步展示了 S-DTM 对于语料库的训练学习结果, 图 4 为映射到“特征抽取”标签和“人工神经网络”标签的两个主题中, 不同时间片上, 出现概率最高的部分词汇。图 5 为映射到“中文信息处理”标签和“语料库”标签的两个主题中, 某些词汇出现的概率随着时间的变化。

从图 4 和 5 中可以得到如下结论。

1) 词汇出现概率最高的词汇与主题标签相关程度高, 说明 S-DTM 的标签约束监督训练取得了成功。另外, 由于标签的存在, 使 S-DTM 训练得到的主题的语义解释变得清晰、明确, 方便理解。

2) 在主题的不同时间片上的词汇概率分布中, 尽管出现概率较高的词汇大致相同, 但是还是有小部分的不同, 这显示了 S-DTM 对于不同时间片上的语言的变化的捕捉精确度较高。并且这个训练结果对于探索某个特定的研究主题在不同历史阶段的研究热点也有较大的帮助。

3) 在主题的词汇分布链上, 同一个词汇的出现概率有较大的不同, 例如, “中文信息处理”主题中, 早期“分词”出现的概率较高, 而后期则呈现下降趋势, 并在近期被词汇“语料库”所超过。而在“语料库”主题中, “系统”, “知识”这两个与主题关系

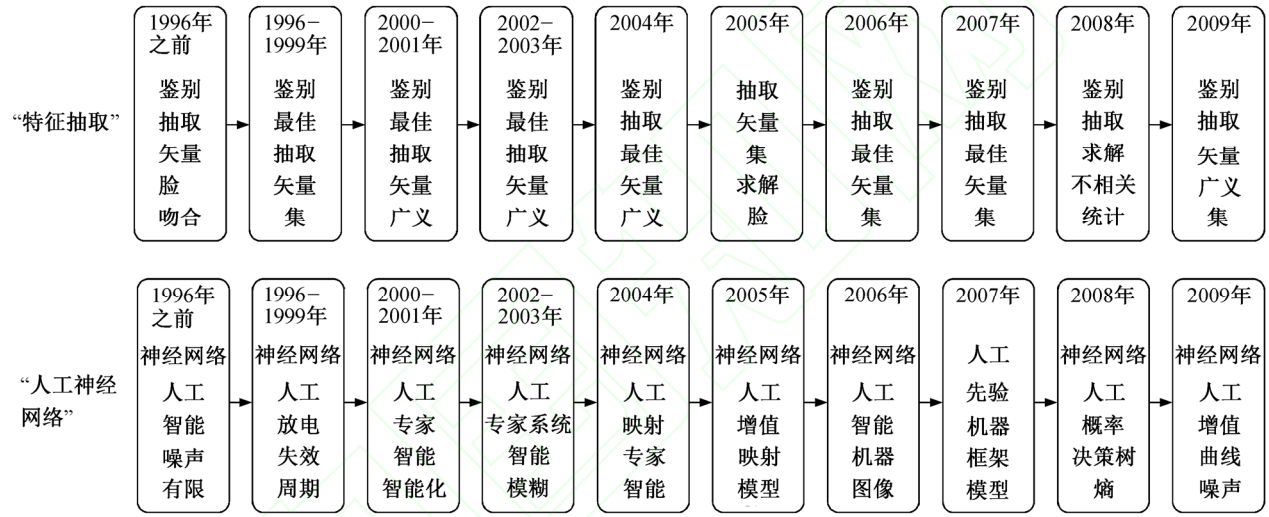

图 4　不同主题上出现概率最高词汇随着时间的变化

Fig. 4　Change of top words appear in two topics over time

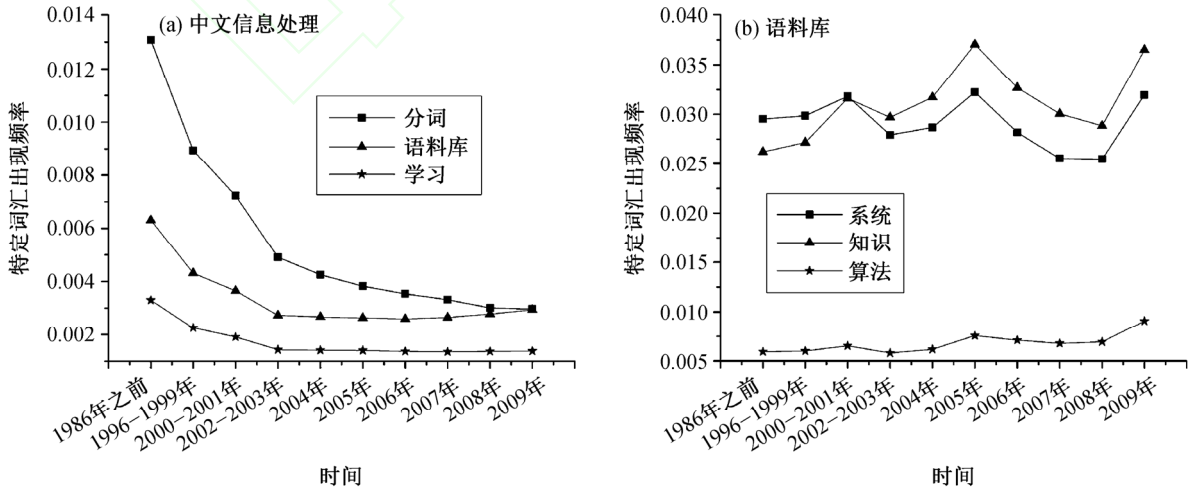

图 5　特定词汇出现概率在不同主题中随着时间的变化

Fig. 5　Probability change of specific words in two topics over time

较为密切的词汇的出现概率随着时间的改变有明显的波动, 而与主题关系不紧密的词汇“算法”的出现概率则较为平稳且保持一个比较低的水平。这充分表明 S-DTM 对语言随着时间的动态变化的建模是成功的。

4 总结

目前概率主题模型存在的主要问题是: 1) 忽视时间要素; 2) 不能对语言动态变化建模; 3) 主题表达解释困难; 4) 主题数目难以确定。针对以上问题, 提出了一个新的结合有监督学习的动态主题模型 S-DTM, 该模型基于 DTM, 结合了 L-LDA 的有监督学习技术, 在模型训练学习中加入主题的标签约束, 克服了传统主题模型存在的不足, 使主题数目的确定变得方便快捷, 主题的表达解释更为清楚, 对于不同时间片上的语言的变化也能够更为精确地捕捉。

本文回顾了当前的相关工作和不足, 并详细介绍了 S-DTM 的生成过程和变分推理过程。通过在一个跨越 25 年“以自然语言处理领域中文期刊论文为主导”的语料库上的一系列实验, 证明了 S-DTM 比静态的 L-LDA 模型和无监督的 DTM 模型在语义概括解释性能、主题结构解释表达性能、主题词汇分布动态捕捉性能上均有所提高。通过 S-DTM 对语料库的学习训练, 可以获得动态的主题–词汇分布, 这将对文本处理、自然语言处理、信息检索等方面的任务, 起到一定的辅助作用。

本文主要存在以下不足: 1) 实验的训练样本数目偏少, 未来需要在拥有更大样本数目的语料库上对 S-DTM 进行验证; 2) 训练时间不够理想, 当主题数目增加时, 训练时间上升明显, 在下一步研究中将会采用在集群上的并行计算来优化模型的训练时间。

参考文献

[1] Blei D M. Probabilistic topic models. Communications of the ACM, 2012, 55(4): 77–84

[2] Blei D M, Carin L, Dunson D. Probabilistic topic models. Signal Processing Magazine, IEEE, 2010, 27(6): 55–65

[3] Blei D M, Lafferty J D. Dynamic topic models // Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 113–120

[4] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2009: 248–256

[5] 中国科学院自动化研究所“自动化学科创新思想与科学方法研究”课题(编号: 2009IM020300)[数据文件]. (2012–01)[2014–06]. http://www.datatang.com/Member/UserData.aspx?id=5878

[6] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. The Journal of machine Learning research, 2003, 3: 993–1022

[7] Wallach H M. Topic modeling: beyond bag-of-words // Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 977–984

[8] Reisinger J, Waters A, Silverthorn B, et al. Spherical topic models // Proceedings of the 27th International Conference on Machine Learning (ICML-10). Israel: Omnipress, 2010: 903–910

[9] Doyle G, Elkan C. Accounting for burstiness in topic models // Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 281–288

[10] Gerrish S, Blei D M. A language-based approach to measuring scholarly impact // Proceedings of the 27th International Conference on Machine Learning (ICML-10). Israel: Omnipress, 2010: 375–382

[11] The Y W, Jordan M I, Beal M J, et al. Hierarchical dirichlet processes. Journal of the American Statistical Association, 2006, 101: 1566–1581

[12] Mcauliffe J D, Blei D M. Supervised topic models // Advances in neural information processing systems. Canada: Neural Information Processing Systems (NIPS), 2008: 121–128

[13] Lacoste-Julien S, Sha F, Jordan M I. DiscLDA: discriminative learning for dimensionality reduction and classification. Advances in neural information processing systems. Canada: Neural Information Processing Systems (NIPS), 2009: 897–904

[14] Winn J. Variational message passing and its applications [D]. Cambridge: University of Cambridge, 2003

[15] Kalman R E. A new approach to linear filtering and prediction problems. Journal of Fluids Engineering, 1960, 82(1): 35–45

[16] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An introduction to variational methods for graphical models. Machine learning, 1999, 37(2): 183–233

[17] mmseg4j 中文分词器. (2014–03)[2014–06]. https://github.com/chenlb/mmseg4j-solr

[18] Mallet. McCallum A K. MALLET: a Machine Learning for Language Toolkit. (2002)[2014–06]. http://mallet.cs.umass.edu

[19] DTM source code. (2011–10)[2014–06]. https://code.google.com/p/princeton-statistical-learning/downloads/detail?name=dtm_release-0.8.tgz

[20] Labeled LDA in Java. (2012–11)[2014–06]. https://github.com/myleott/JGibbLabeledLDA

[21] Griffiths, Thomas L, Steyvers M. Finding scientific topics // Proceedings of the National academy of Sciences of the United States of America 101. 2004: 5228–5235