文章编号：1003-0077(2010)06-0043-07

基于 LDA 话题演化研究方法综述

单 斌，李 芳

（上海交通大学 计算机科学与工程系 中德语言技术联合实验室，上海 200240）

摘 要：现实生活中不断有新话题的产生和旧话题的衰减，同时话题的内容也会随着时间发生变化。自动探测话题随时间的演化越来越受到人们的关注。Latent Dirichlet Allocation 模型是近年提出的概率话题模型，已经在话题演化领域得到较为广泛的应用。该文提出了话题演化的两个方面：内容演化和强度演化，总结了基于 LDA 话题模型的话题演化方法，根据引入时间的不同方式将目前的研究方法分为三类：将时间信息结合到 LDA 模型、对文本集合后离散和先离散方法。在详细叙述这三种方法的基础上，针对时间粒度、是否在线等多个特征进行了对比，并且简要描述了目前广泛应用的话题演化评测方法。文章最后分析了目前存在的挑战，并且对该研究方向进行了展望。

关键词：话题模型；话题演化；Latent Dirichlet Allocation

中图分类号：TP391　　文献标识码：A

A Survey of Topic Evolution Based on LDA

SHAN Bin, LI Fang

(Sino-German Joint research Lab for Language Technologies,
Dept. of Computer Science & Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: With topics evolve over time, new topics emerge and old ones decay. Many researches are devoted to detect the topic evolution automatically. Latent Dirichlet Allocation (LDA), as a recently emerged probabilistic topic model, has been widely used in the research of topic evolution. This paper discusses two aspects of evolution on topic, i.e. the content and the topic intensity. It summarizes three methods in LDA based topic evolution detection according to the dealing with time: joining the time to LDA model, post-discretizing or pre-discretizing methods. The three methods are also compared in several features: the time granularity, on-line or off-line, etc. In addition, the evaluation methods for topic evolution are introduced. Finally, the paper gives some analysis and suggestions for future researches on topic evolution based on LDA.

Key words: topic model; topic evolution; Latent Dirichlet Allocation

1 引言

互联网已经成为人们获取信息的一个主要渠道，突发新闻事件或新闻话题可以在互联网上瞬间传播，如何跟踪该新闻话题或新闻事件的后续发展，是人们关心和需要迫切解决的问题。随着时间的发展，新闻话题的内容会发生变化，新闻话题的强度也会经历一个从高潮到低潮的过程。如何有效地组织这些大规模文档，并且按时间顺序来获取文本集合中话题的演化，从而帮助用户追踪感兴趣的话题，具有实际意义。更重要的是，在新闻专题报道和一些安全机构针对犯罪探测和预防的任务中，更需要从文本集合中快速准确地追踪话题的演化并且根据演化做出相应的预测。因此，话题演化研究具有现实的应用背景。

收稿日期：2009-12-02　定稿日期：2010-04-01

基金项目：国家自然科学基金资助项目(60873134)

作者简介：单斌(1986—)，男，硕士，主要研究领域为自然语言处理，信息检索和信息抽取；李芳(1963—)，女，博士，副教授，主要研究领域为自然语言处理，信息检索与信息抽取。

早在话题检测与跟踪(Topic Detection and Tracking,简称TDT)研究中,人们就已经认识到对新话题的自动识别和已知话题的持续跟踪的重要性。在TDT中,话题被定义为一个种子事件或活动以及与之相关的所有事件或活动[1]。话题跟踪(Topic Tracking)主要就是跟踪已知话题的后续报道,采用相似度计算公式来判断新话题是否属于已知话题,主要方法基于统计知识,对文本进行信息过滤,然后利用分类策略来跟踪相关话题,但是TDT早期的研究并没有有效利用语料的时间信息,在时间轴上分析话题的分布。

随着话题模型[2-4]的兴起,如何借助话题模型,引入文本语料的时间信息,研究话题随时间的演化,成为在机器学习领域、文本挖掘领域研究的热点。不同于TDT中话题的表示,话题模型假设:每篇文本是话题的混合分布,而每一个话题是一组词语的混合分布[5]。话题模型借助话题可以很好地模拟文本的生成过程,对文本的预测也有很好的效果,因此在话题演化领域有着一定的优势,目前关于这方面的研究已经有很多方法和成果[6-11]。

本文将主要关注基于Latent Dirichlet Allocation(简称LDA)话题模型[3]的话题演化方法。首先简要的介绍LDA技术以及相关概念,第3部分着重介绍各种基于LDA的话题演化方法,第4部分对所有方法进行总结比较,第5部分介绍话题演化的评测方法,最后,对全文进行总结,并对该研究方向进行展望。

2 基本概念介绍

2.1 LDA话题定义以及模型简介

在话题演化研究中,一个重要的任务就是获取文本集合的话题。话题实际就是文本的一种降维表示。最早的文本降维技术是词频—反文档频率(Term Frequency-Inverse Document Frequency,简称tf-idf),但tf-idf无法在语义层面表示文本。随后Deerwester等人[12]利用矩阵的奇异值分解技术对文本降维,即隐性语义索引(Latent Semantic Indexing,简称LSI)模型。Hofmann[2]在LSI基础上提出了概率隐性语义索引模型(probabilistic Latent Semantic Indexing,简称pLSI),它假设每篇文档是由多项式随机变量(话题)混合而成,而文档中每个词,由一个话题产生,文档中不同的词可有不同的话题生成。但是pLSI模型参数数量随着文集增长而线性增长,并且会产生过拟合的问题。

Blei等人[3]在2003年提出了Latent Dirichlet Allocation(简称LDA)模型。LDA模型是一个概率生成模型,同时也是一个话题模型,它的参数不会随着文集增长而线性增长,有很好的泛化能力,是机器学习、信息检索等领域很流行的一个模型。目前,为满足不同的需求,出现了很多基于LDA的扩展模型和应用模型,例如文献[13-14]。

下面先介绍LDA模型中使用的符号,见表1。

表1 文中用到的符号

符号	符号的描述
D	文集中文档的集合
K	话题的集合
N_d	文档d的长度(词语的个数)
$w_{d,i}$	文档d中的第i个词
$z_{d,i}$	文档d中的第i个词的话题
α	LDA模型的*Dirichlet*先验分布,表示整个文集上话题分布的先验
β	LDA模型的*Dirichlet*先验分布,表示所有话题上词语分布的先验
θ_d	表示文档d上话题的多项式分布
φ_z	表示话题z上词语的多项式分布

LDA是三层的变参数层次贝叶斯模型,假设一篇文档是由一些潜在的话题的多项式分布表示,而话题由一组词的多项式分布组成。所以又叫话题模型。模型描述了文档的生成过程,步骤如下:

1)对于每个文档$d\in D$,根据$\theta_d \sim Dir(\alpha)$,得到多项式分布参数$\theta_d$;

2)对于每个话题$z\in K$,根据$\varphi_z \sim Dir(\beta)$,得到多项式分布参数$\varphi_z$;

3)对文档d中的第i个词$w_{d,i}$

a)根据多项式分布$z_{d,i} \sim Mult(\theta_d)$,得到话题$z_{d,i}$。

b)根据多项式分布$w_{d,i} \sim Mult(\varphi_z)$,得到词$w_{d,i}$。

在LDA中,话题(Topic)由一组语义上相关的词语以及词语在该话题上出现的概率表示。即:话题$z=\{(w_1, p(w_1|z)), \cdots, (w_V, p(w_V|z))\}$,其中$p(w_v|z)$表示已观测到话题$z$的情况下词语$w_v$出现的概率。

2.2 话题演化

话题演化衡量的是同一话题随时间推移表现出的动态性、发展性和差异性。话题的演化定义为话题随时间的变化，而这个变化往往反应在两方面，第一，就是话题强度随着时间推移发生的变化，例如，四年一届的奥运会，在奥运年受关注度高，而在非奥运年，受关注度低。第二，就是话题内容随着时间的推移而发生的变化，具体到基于 LDA 的话题，就是表示话题的词语和词语的分布概率的变化。例如：在奥运会前夕，大家关注奥运会的准备工作，奥运会结束后，大家关注对奥运会的总结和盘点。话题强度的演化衡量的是话题受关注程度的变化，话题内容的演化衡量的是话题关注点的迁移，从而体现出话题的动态性、发展性和差异性。

3 基于 LDA 的话题演化方法介绍

目前基于 LDA 的话题演化方法，在内容演化和强度演化上有各自不同的特点。根据引入时间方式的不同，我们总结了三种不同的演化方法：第一种方法是将时间作为可观测变量结合到 LDA 模型中；第二种方法是在整个文本集合上用 LDA 模型生成话题，然后按文本的时间信息，根据话题后验离散地分析话题随时间的演化；第三种方法将文本集合先按一定时间粒度离散到相应的时间窗口，在每个窗口上运用 LDA 模型来获取演化。下面依此对上述三种方法进行详细阐述。

3.1 将时间信息结合(Joint)到 LDA 模型中

这种方法将文本的时间信息作为可观测变量，结合到 LDA 话题模型中，指导文本集合上话题的分布，这样，话题表现出在时间轴上强度的演化。

基于这种方法的代表模型是 Topic Over Time (简称 TOT)模型[6]。TOT 模型不依赖于马尔科夫假设，而是将时间看作连续的可观测变量。TOT 模型假设每个词的生成不仅仅受到它所属的话题的限制，同时也受到时间属性的影响，因此可以更好的描述每个话题在不同时间窗口的强度。

TOT 的模型生成过程与 LDA 模型类似，只是每个词语 $w_{d,i}$ 多了一个时间属性 $t_{d,i}$，而 $t_{d,i}$ 由连续贝塔(Beta)分布($t_{d,i} \sim Beta(\psi_{d,i})$)生成，其中 $\psi_{d,i}$ 为文档 d 中词语 i 的时间先验分布。虽然同 LDA 模型一样，话题内容是不变的，但是由于 TOT 模型考虑了文本的时间信息，所以可以表示话题在不同时刻的分布强度，使得 TOT 模型生成的话题比原始 LDA 模型生成的话题在时间分布上更准确，也具有更好的可解释性。

TOT 模型的优点是模型的时间是连续的，不会出现在离散时间的方法中时间粒度选取的问题，而在很多语料中，时间粒度的选取决定了最后结果的好坏。但是 TOT 模型所展示的话题在时间上的演化，仅仅是指话题强度的变化趋势，而忽略了话题内容的变化。另外，TOT 是基于 LDA 模型的改进，所以 TOT 是离线的对文集进行处理，不具备扩展性，必须一次对所有的文档运用 TOT 模型。对于新观测到的文本，必须重新建模。

3.2 后离散分析(Post-discretized Analysis)

这种方法是在先忽略时间的情况下，在整个文本集合上运用 LDA 或者 LDA 的改进模型获取话题，然后利用文本的时间信息检查话题在离散时间上的分布来衡量演化，称为后离散分析(Post-discretized Analysis)方法。

Griffiths 等人[15]在 2004 年首先提出了这种方法。先在整个文集上用 LDA 话题模型获取所有的话题(文献[15]提出了用模型选择的方法确定话题的个数)，进而估计出 LDA 模型的参数(θ_d, φ_z, z)。然后按照文档的发布时间，将文档离散到相应的时间窗口。对于某个话题 z_k，依次考虑它在每个时间窗口的强度 δ。则：

$$\delta = \frac{1}{D_t} \sum_{d;\, t_d \in t} \theta_{dk} \tag{1}$$

D_t 表示属于时间窗口 t 的文档数量。

从而显示了随时间推移，强度明显上升的热话题(hot topic)和下降的冷话题(cold topic)。

另一种后离散分析的方法由 Hall 等人[7]在 2008 年提出，通过计算话题在以年为粒度的离散时间上分布的后验概率来表示话题分布的强度。

$$\begin{aligned} p(z \mid y) &= \sum_{d;\, t_d = y} p(z \mid d) p(d \mid y) \\ &= \frac{1}{C} \sum_{d;\, t_d = y} \sum_{z_i \in d} I(z_i = z) \end{aligned} \tag{2}$$

其中每篇文档仅属于一个时间窗口 t_d，且 $P(d \mid y)$是一个常量 $1/C$，表示文档 d 在时间窗口 y 上出现的概率，$P(z \mid d)$表示话题 z 在文档 d 上出现的概率，由该话题在文档上出现次数的指示函数计算而来。这种方法很好地衡量了科学领域话题发展的趋势。

以上两种方法都主要应用于追踪科学领域的话题强度演化，实验文集也都来自于科学领域的会议或期刊。这是因为会议期刊的时间粒度是确定的（文献[7]中实验数据来自ACL，COLING会议，他们是每年举行一次），而且每篇文章的内容具有差异性，同时每一年发表的文章一定基于前几年的研究结果，这保证了话题演化的特性。与TOT模型相同，它们衡量的话题演化是基于话题强度，而不是基于话题内容的演化；另外这种方法也是基于在整个文集上一次性获取话题，所以是离线的，很难扩展到基于流的数据集。不过比起下文提及的先离散再获取话题（pre-discretized）的方法，post-discretized方法没有话题对齐（align）的问题。但是很明显，这种post-discretized的方法依赖于话题在时间上分布的后验的计算方式，两种方法对于强度的具体计算公式不同，但是，表现的意义确是相似的。

3.3 按时间先离散（Pre-discretized）方法

文本先根据其时间信息离散到时间序列上对应的时间窗口内，然后依次地处理每个时间窗口上的文本集合，最终形成话题随时间的演化，因此被称为先离散（pre-discretized）分析的方法。

先离散方法有各自不同的特点。从处理文集的类型上：有的模型处理的是封闭的文本集合，如文献[8]；有的处理基于流的数据集合，如文献[17]。从演化的时间粒度上：很多模型的时间粒度，往往受限于模型处理的文本集合，有的以年为粒度，如文献[8]，有的可以以天为粒度，如文献[9]，有的模型可以从不同的时间粒度展现话题演化，如文献[16]，还有的模型基于连续的时间，如文献[19]。

另外，在先离散分析的方法中，下一时刻的模型参数往往依赖于当前时刻（或前几个时刻）的模型参数的后验或者模型输出结果。这种依赖表现为条件概率依赖[22]或者非条件概率依赖。本节将从这个角度详细介绍这两种基于先离散方法的模型。

3.3.1 基于条件概率的先离散方法

这种方法的主要思想是当前时刻的模型参数后验作为下一时刻模型参数的条件分布引入模型，这样从全局上看，整个话题演化模型依然是图形模型（Graphic Model），但在模型参数推导过程中可能比较困难。另外对全局的处理使得通过一次建模就可以得到所有时刻的话题表示，但不具有在线添加新文本的功能，对于新到达的文本只能重新离散、全局建模。

这种方法的代表之一就是动态话题模型（Dynamic Topic Model，简称DTM）[8]。DTM先根据时间窗口分割文本集合，并假设话题数量K是固定的，即每个时间窗口的文本都由K个话题的LDA模型生成。

DTM用状态空间模型来实现演化。在DTM中，实际获取的演化特征是话题在文集上分布的演化以及词语在话题上的分布的演化，即话题的分布强度和话题的内容都在随着时间而演化。

由于DTM将时间离散，所以演化的效果决定于时间粒度的选择，粒度太大会导致演化并不真实，粒度太小使得在模型参数推导中引入太多的时间节点。为了解决DTM中时间粒度的问题，Chong Wang等人提出了连续时间的动态话题模型（Continuous Time Dynamic Topic Model，简称为CTDTM）[19]。CTDTM用布朗运动（Brownian Motion）模型来实现话题的演化过程，并将文本的时间差信息引入到参数演化的过程中，可以看作是选取最佳时间粒度下的DTM模型。所以，无论是DTM还是CTDTM，在获取演化的能力上，是类似的。

另一种基于条件概率的先离散方法是动态混合模型（Dynamic Mixture Model，简称DMM）[17]。DMM与DTM（或CTDTM）相比，具有更强的时间假设。在DMM中的文本是严格按照时间顺序先后到达的，每个时刻只到达一篇文本，从这个角度DMM可以看作在线的话题演化模型。DMM假设模型参数θ由前一时刻θ_{t-1}的混合分布生成。即：

$$\theta_{t+1} \mid \theta_t \sim Dir(\varphi\theta_t) \quad (4)$$

从DMM的演化依赖关系上，说明了DMM假设连续两篇文档中话题的分布存在演化关系，所以更适用于获取文本间更细微的内容和强度的演化。

Multiscale Topic Tomography模型（简称MTTM）[16]也是基于这种方法的模型。但与前面的模型不同，MTTM更关注于多时间粒度的话题演化。MTTM用泊松过程来模拟文档的生成，用泊松参数来表示词语在话题上出现的期望次数。MTTM把时间重复的分割成相等地两个时间窗口，最终时间窗口形成二叉树的层次结构，进而假设父时间窗口上模型的泊松参数由其左右孩子时间窗口的泊松分布按一定比例组合成。

经过参数推导简化后，可以估计出不同粒度上的模型参数，也就可以表示话题内容和强度的演化。因而MTTM模型不仅体现出TOT模型衡量话题演化强度的性质，也体现出DTM模型衡量话题内

容演化的性质。

3.3.2　基于非条件概率的先离散方法

基于非条件概率的方法中，当前时刻的模型参数后验或输出结果直接用来计算下一时刻的模型参数，而不存在条件依赖的关系，这样虽然每个时刻模型依然是图形模型，但是从全局上看并不是一个图形模型。非条件概率依赖的好处是：保持了 Dirichlet 先验分布，从而使得模型的参数推导非常方便，而且由于独立获取每个时间窗口的话题，使得模型具有在线处理的能力，对于新到达的文本(或文本集合)可以增量处理。

据我们所知，最早提出对 LDA 模型按照文本达到时间来增量建模的方法是增量 LDA (Incremental Latent Dirichlet Allocation，简称 ILDA)[18] 算法。ILDA 算法利用了 T. L Griffiths 和 M. Steyvers[15] 中提出的用 Gibbs 采样方法，估计 LDA 的话题后验分布和 LDA 模型参数。其中每个时间段上的话题个数，都由独立的贝叶斯模型选择方法来确定，因此 ILDA 的演化话题个数是可变的。ILDA 算法获取的演化是话题上词语分布的演化，展现出话题内容的变化。

另一个方法是 L. AlSumait 在 2008 年提出的在线 LDA 模型(Online Latent Dirichlet Allocation，简称 OLDA)[9]，OLDA 模型在线地生成一个及时更新的模型，可以表现出话题内容和强度的演化过程。OLDA 使用演化矩阵来记录以前的模型结果，而且利用演化矩阵实时地检测新话题的产生。OLDA 的基本思想是，通过处理新到达的文本，调整学习到的话题，同时利用演化矩阵 B^{t-1} 记录话题的变化，这样就可以丢弃前面处理过的所有文本。演化矩阵 B_k^t 表示了话题 k 随时间在内容上的演化。另外，可以通过计算演化矩阵连续两列的 Kullback-Leibler 差分距离，度量在连续时间窗口内，同一话题在内容上的差异性。如果 KL 差分距离大于阈值，就认为探测到新话题。

OLDA 模型不像 DMM 模型按严格的时间顺序依次处理文本，同样也不像 DTM 模型需要一次处理较大的文本集，OLDA 模型的时间粒度可以介于 DMM 和 DTM 模型之间。更深入地，L. AlSumait 等人[21] 详细地分析了关于 OLDA 中演化矩阵时间窗口大小 δ和权重的 ω 的选择方法，使得 OLDA 展现出更好的效果。

4　基于 LDA 话题演化方法的比较

本节主要对第三部分提到的各种模型方法进行总结比较，见表 2。根据话题演化任务关注的特征，我们选择了是否在线，引入时间方式，时间粒度，话题数量等特征来比较。是否在线主要考察模型对于新观测文本的处理能力；引入时间的方式和时间粒

表 2　基于 LDA 话题演化方法比较

	代表模型	是否在线	引入时间方式	时间粒度	演化类型	话题数量
Joint	TOT	否(基于有限文本)	作为可观测连续变量	连续，不存在粒度问题	强度演化	固定
Post-discretized	Scientific Topic	否(基于有限文本)	按时间后离散	需要较大粒度(如：年)	强度演化	固定
	Historical Idea	否(基于有限文本)	按时间后离散	需要较大粒度(如：年)	强度演化	固定
Pre-discretized	DTM	否	先按时间离散	需要较大粒度(如：年)	内容和强度演化	固定
	CTDTM	否	连续时间	自由选择	内容和强度演化	固定
	DMM	是(新文本要重新建模)	严格按文本时间顺序处理	粒度较小	内容和强度演化(微观)	固定(但数目较少)
	MTTM	否(基于有限文本)	先按时间离散	多时间粒度	内容和强度演化	固定
	OLDA	是	按文本时间顺序处理	粒度适中	内容和强度演化	固定(可以检测新话题)
	ILDA	是	先按时间离散	粒度适中	内容演化	数目可变

度的选择说明了模型获取演化的细致程度;演化类型主要表明模型在强度演化和内容演化两方面的能力;最后话题数目主要说明模型对新话题或衰亡话题的探测。

5 话题演化的评测方法

话题演化任务中,有一些常用的评测指标:话题的相似度、模型的泛化能力,以及演化结果的评测上。需要指出的是,在演化结果的评测上,目前并没有统一的标准。

5.1 话题相似度度量

首先,在话题演化任务中,有一些方法中话题并不是对齐的[10],所以需要一些衡量话题相似度的方法来对齐话题。另外,即使有些话题模型的话题是对齐的,但是往往为了探测话题的产生[9],同样需要衡量话题相似度。

采用比较多的话题相似度度量方法,是利用Kullback-Leibler差分距离的方法[5]。话题j_1和话题j_2的不相似度,由对称的Kullback-Leibler距离衡量,即:

$$KL(j_1, j_2) = \frac{1}{2}\sum_{k=1}^{W}\varphi_k^{(j_1)}\log_2\varphi_k^{(j_1)}/\varphi_k^{(j_2)} + \frac{1}{2}\sum_{k=1}^{W}\varphi_k^{(j_2)}\log_2\varphi_k^{(j_2)}/\varphi_k^{(j_1)} \quad (9)$$

其中φ表示话题—词语的分布参数。根据具体实验,设定合适的阈值,可以判断话题是否同一或者是否新生。

当然其他的一些距离度量方法也可用于话题相似度的衡量,如余弦距离,Jenson-Shannon距离等。

5.2 模型泛化能力度量

模型泛化能力是衡量模型对于未观测到的数据的预测能力。比较公认的判断方法是衡量模型的困惑度(Perplexity)。模型的困惑度往往与基础(Baseline)模型的困惑度进行对比,来说明新模型对于预测未观测数据有更好的能力。困惑度表示为:

$$Perplexity(D_{test}|\mathrm{M}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \quad (10)$$

困惑度越小,表示模型的泛化能力越强。

5.3 演化结果评测

5.3.1 话题内容演化评测

某一时间段话题内容的表示是一组词,也就是用模型在每个时间窗口t上的话题—词语分布的后验参数$\beta_{t,z}$或$\varphi_{t,z}$,按照概率来排序,将出现频率最高的W个词语来显示话题。话题内容演化是否正确是根据人的判断。

另外,话题在内容上的变化也可以用同一个词语(word)在同一个话题上出现的次数(即概率)随时间的变化来表示。

5.3.2 话题强度演化评测

话题强度的演化图,用坐标图来表示,一般横轴表示时间,纵轴表示话题k的概率,即$p(k)$,这个概率可以由模型的后验来得到。

在先离散的方法中,由于文本已经被划分到相应的时间窗口,所以可以直接利用时间窗口t中的每个文档的参数$\theta_{t,d}=\{\theta_{t,d,1}, \cdots, \theta_{t,d,K}\}$(文档上话题的分布概率)的平均值来计算每个话题在$t$时刻的出现强度,以此来衡量话题强度的演化。

在后离散和引入时间观测变量的方法中,要计算话题k在时间t上的后验$p(k|t)$来得到话题的强度分布。一般来说,这要借助于文本的时间信息。

目前,话题内容演化和话题强度演化没有统一的评判标准,也没有有效的量化比较,只是通过人工来自动判断。上述这些评测方法是根据许多参考文献总结的。

6 存在问题的分析以及展望

本文详细介绍了基于LDA话题模型的话题演化各种不同的方法。按照引入时间的方式,将基于LDA的话题演化技术分为:直接把时间作为观测变量引入模型、按时间后离散和先离散三种方法。直接将时间引入模型,可以自然地探测到话题强度的变化,无须考虑时间的粒度。后离散方法简单,基于静态的词语集合和话题数目,不易扩展。先离散方法因为符合人们观测文本信息的事实,受到更多关注。根据演化的不同特征,我们对比和总结了各种不同的方法,见表2,不同的方法有其不同的特点,可以应用在不同的任务中。

但是基于LDA的话题演化课题依然处于研究阶段。作者认为主要是该课题中还有很多需要解决的问题和技术难点。

首先，从我们的对比中可以看出，大多数基于 LDA 话题演化方法都假设话题数目是固定的，无法探测新话题的产生，旧话题的消亡和分裂，这不符合现实中的话题。如果假设不同时间段话题数目不同，那么这涉及到如何定义同一话题和相关话题，在话题模型里，如何定义和区分同一话题和相关话题具有一定的挑战。目前的大多数研究都回避了这一问题，通过假设话题数目固定，不同时间段话题对齐，忽略了话题的消亡、分裂、迁移的可能。因此在今后的研究中，需要提出一种新的方法和明确的定义，来判断同一话题或者相关话题，从而发现随时间的话题演化关系。

其次，随着 LDA 模型的广泛使用，对 LDA 话题的表示和话题可解释性问题备受学者的关注。很多学者致力于这方面的研究，有一些初步的结果。这些研究中，一种是基于 LDA 的扩展模型，通过引入其他的特征指导 LDA 话题生成，例如引入文章的作者信息（Author-Topic Model）[14]、科学研究论文中的参考文献信息（Citation LDA）[23] 等；另一种是通过半监督或监督的方法指导 LDA 话题的生成，如 Supervised Topic Model[13]；对话题的表示，最近的研究是通过对表示话题的词语进行分析组合，用更有意义的词组（n-gram）代替单个词语来表示话题，如文献[24]。因此，如何把自然语言处理技术以及其他的技术引入到基于 LDA 模型的话题演化任务中，构造更明确清晰的话题演化，也将是我们面临的又一项挑战。

最后，虽然已经有很多关于话题演化的研究，但是对于话题演化的评测，没有一个评判标准，没有统一的测试指标和相应的测试语料。目前，无论是话题强度还是内容的演化，都是基于人们对话题的主观理解。这种方法不具有可比性，对于话题未来的发展趋势的预测也不是很科学。所以，提出一个话题演化的评判标准也是需要解决的问题之一。

尽管基于 LDA 话题演化研究存在着众多挑战，但是，LDA 话题模型能够自动获取海量文本信息的主题或话题，它是一种非监督的方法，具有实际应用的前景，因此，基于 LDA 的话题演化研究仍然受到很多关注，发展也很快，我们相信，随着研究的深入，问题的解决，话题演化研究一定会得到广泛的实际应用。

参考文献

[1] 洪宇，张宇，刘挺，等. 话题检测与跟踪的评测与研究综述[J]. 中文信息学报，2007，21(6)：71-87.

[2] Thomas Hofmann. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley，CA，USA，1999：50-57.

[3] David M. Blei，Andrew Y. Ng，Michael I. Jordan. Latent dirichlet allocation[J]. The Journal of Machine Learning Research，2003，3：993-1022.

[4] T. Griffiths，M. Steyvers. A probabilistic approach to semantic representation[C]//Proceedings of the 24th Annual Conference of the Cognitive Science Society. Mahwah，NJ：Erlbaum，2002，381-386.

[5] M. Steyvers，T. Griffiths. Probabilistic topic models. In：T. Landauer，D. S. McNamara，S. Dennis，W. Kintsch（Eds.），handbook of Latent Semantic Analysis[M]. Hillsdale，NJ.. Erlbaum. 2007.

[6] X. Wang，A. McCallum. Topic 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. Philadelphia，PA，USA，2006：424-433.

[7] D. Hall，D. Jurafsky，C. D. Manning. Studying the history of ideas using topic models[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu，Hawaii，2008：363-371.

[8] D. M. Blei，J. D. Lafferty. Dynamic topic model[C]//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh，Pennsylvania，2006：113-120.

[9] L. Alsumait，D. Barbara，C. Domeniconi. On-line LDA：Adaptive topic models of mining text streams with applications to topic detection and tracking[C]//Proceeding of the 8th IEEE International Conference on Data Mining. Washington，DC，USA：IEEE Computer Society，2008：3-12.

[10] 楚克明. 基于 LDA 新闻话题的演化[C]//第五届全国信息检索学术会议. 上海，中国，2009：64-72.

[11] A. Gohr，A. Hinnerburg，R. Schult，M. Spiliopoulou. Topic evolution in a stream of documents[C]//Proceeding of the Society for Industrial and Applied Mathematics. 2009：859-870.

[12] S. Deerwester，S. Dumais，T. Landauer，etc. Indexing by latent semantic analysis[J]. Journal of the American Society of Information Science，1990，41(6)：391-407.

[13] D. M. Blei，J. D. McAuliffe. Supervised topic models[C]//Proceeding of the 22nd Annual Conference on Neural Information Processing Systems，2008.

[14] M. Rosen-Zvi，T. Griffiths，M. Steyvers，etc. The

（下转第 68 页）

5 展望

混合高斯模型能够较好地刻画说话人的特性，当高斯混合数越多的情况下，刻画也就越准确。本实验采取的是单高斯模型，所以对目标说话人区分能力较弱，因此一个直观的想法就是增加混合数目。但是，对于嵌入式应用中训练遍数不多的情形，这必然引起训练数据不足的问题。一个解决的方法是引入通用背景模型（Universal Background Model，UBM）并利用自适应算法进行说话人模型训练。当然，引进UBM势必会增加系统的时间开销，因此我们计划在未来的研究中，UBM的训练脱机完成，说话人模型的训练采用快速自适应算法并仅取最大似然的前几个高斯混合，这可以大大降低建模过程的时间消耗，而不至于影响说话人模型的描述精度。

参考文献

[1] 蒋力. 基于概率统计模型的非特定人语音识别方法与系统的研究[D]. 北京：清华大学，1989.11.

[2] 郑方. 非特定人连续数字识别方法与汉语语音数据库的研究[D]. 北京：清华大学，1992.

[3] 郑方，吴文虎，方棣棠. CDCPM及其在语音识别中的应用[J]. 软件学报，1996，7：69-75.

[4] Thomas Fang Zheng, Chai Haixin, Shi Zhijie. A real-world speech recognition system based on CDCPMs [C]//Int'l Conf. on Computer Processing of Oriental Languages (ICCPOL'97), 1997, 1: 204-207.

[5] Reynolds A. Reynolds. Speaker identification and verification using Guassian mixture speaker models[C]// Speech Communication. 1995, 17(1-2): 91-108.

[6] N. Z. Tisby. On the application of mixture AR hidden Markov models to text independent speaker recognition [C]//IEEE Trans. Signal Processing, March 1991, 39(3): 563-570.

[7] Zheng, F., Wu, W.-H., Fang, L.-T., A Log-Index Weight Cepstral Distance Measure for Speech Recognition [J]. J. of Computer Science and Technology (JCST), 1997, 12(2): 177-184.

（上接第49页）

author-topic model for authors and documents[C]// Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada, 2004: 487-494.

[15] T. L. Griffiths, M. Steyvers. Finding scientific topics [C]//Proceeding of the National Academy of Science of United States of America, 2004, 101: 5228-5235.

[16] R. M. Nallapati, S. Ditmore, J. D. Lafferty, etc. Multiscale topic tomography[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA, 2007: 520-529.

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

[18] X. Song, C. Y. Lin, B. L. Tseng, etc. Modeling and predicting personal information dissemination behavior[C]//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, Illinois, USA, 2005: 479-488.

[19] C. Wang, D. Blei, D. Heckerman. Continuous time dynamic topic models[C]//Proceeding of the 23rd Conference on Uncertainty in Artificial Intelligence, 2008.

[20] D. M. Blei, J. D. Lafferty. Correlated topic model [C]//Advances in Neural Information Processing System 17. Cambridge, MA: MIT Press, 2005.

[21] L. AlSumait, D. Barbara, C. Domeniconi. The role of semantic history on online generative topic modeling [R]. http://www.ise.gmu.edu/~carlotta/publications/Siam_SemOLDA.pdf: 2009.

[22] G. Shafer. Advances in the understanding and use of conditional independence[J]. Annals of Mathematics and Artificial Intelligence, 1997, 21(1): 1-11.

[23] R. Nallapati, A. Ahmed, E. P. Xing, etc. Joint latent topic models for text and citations[C]//Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, Nevada, USA, 2008: 542-550.

[24] D. M. Blei, J. D. Lafferty. Visualizing topics with multi-word expressions[J]. The Journal of Machine Learning Research, 2009, 7.