基于 LDA 的主题演化研究*

李 勇 安新颖

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 通过监测主题在不同时间窗口内的变化趋势进行主题演化分析，在各时间窗口中分别建立 LDA 模型，采用 Gibbs 抽样方法求解 LDA 模型中的潜在变量，利用 Kullback - Leibler 距离来衡量主题之间的相似度，利用改进的 Z - Score 方法计算主题随时间的偏移程度以反映其演化情况。

〔关键词〕 主题模型；演化；主题偏移

Researh on Topic Evolution Based on LDA *LI Yong, AN Xin - ying, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 In the paper topic evolution analysis is achieved by tracking the topic trends in different time - slices. Latent Dirichlet Allocation (LDA) model is built in time - slices. Gibbs algorithm is used to find out latent variables in LDA modle, Kullback - Leibler divergence is used to measure the similarity between topics. the modified Z - score method is used to measure the drift between topics in order to reflect topic evolution.

〔**Keywords**〕 Topic model; Evolution; Topic drift

1 引言

1.1 主题演化和 LDA 模型

主题演化是主题随时间的变化，包括两方面内容：（1）主题强度随着时间推移发生变化。（2）主题内容随时间的推移发生变化。随着时间的推移，研究主题的内容会发生变化，其受关注的程度也会经历一个从高潮到低潮的变化过程。如何跟踪研究主题的后续发展，是研究人员关心和迫切需要解决的问题。

主题实际上是文本的一种降维表示。最具代表性的文本降维技术是 tf - idf。随后 Deerwester 提出了潜在语义索引（Latent Semantic Indexing, LSI）模型[1]，其主要思想是基于矩阵的奇异值分解技术，对文本进行降维。随后，Hofmann 在 LSI 的基础上又提出了概率潜在语义索引模型（Probablilistic Latent Semantic Indexing, pLSI），该模型假设文档由多个主题混合而成，而文档中每个词由一个主题按一定概念产生，每个词也可由不同的主题生成。与此同时该模型的参数数量随着文集增长而呈线性增长趋势，即出现过拟合问题[2]。2003 年 Bleii 等提出了潜在狄利克雷分配模型（Latent Dirichlet Allocation, LDA）[3]，该模型是一种概率生成模型，使用 Dirichlet 分布描述文档的主题混合比例，模拟文档产生过程。

〔收稿日期〕 2012 - 11 - 27

〔作者简介〕 李勇，助理研究员，发表论文 10 余篇。

〔基金项目〕 国家科技支撑计划课题“基于 STKOS 的科技监测应用示范”（项目编号：2011BAH10B06 - 02）；教育部人文社会科学研究项目“基于知识组织体系的科技文献新主题监测研究”（项目编号：11YJC870001）。

万方数据

1.2 相关研究

目前已有许多研究利用 LDA 模型开展主题演化分析，其基本思路是分析主题和词的概率分布随时间的变化情况，包括主题受关注程度的变化以及受关注点的迁移。根据引入时间的不同，基于 LDA 的主题演化方法可总结为 3 种[4]：（1）将时间作为可观测变量结合到 LDA 模型中。（2）在全部文档集合中应用 LDA 模型生成主题，然后根据文档的时间信息分析主题随时间的演化。（3）首先将文档集合划分到相应时间窗口，然后使用 LDA 模型来进行演化分析。在本文中将采用先离散后刻画主题的方法。

2 主题演化

2.1 基于 LDA 的主题演化方法

生物学中对演化的定义为生物在不同世代之间具有差异的现象，是由于生物的可遗传变异以及生物对环境的适应和物种间的竞争。经过大自然的选择，物种的特征被保留或者淘汰，从而导致新物种诞生或者原物种消失，演化是推动自然发展的动力。参照生物学中的定义，主题演化分析就是从连续文档流中抽取主题并发现其演化规律的过程。本文采用主题 - 词概率密度来表示主题。主题演化就转变成主题 - 词概率分布随着时间的变化产生偏移。在本文中，将采用按时间先离散的方法，把文本流按一定粒度划分成多个时间窗口，然后对每个时间窗口内的文档采用 LDA 模型进行建模，得到文档的主题 - 词概率分布和文档 - 主题概率分布。然后使用 KL 距离计算主题 - 词概率分布的差异，并使用改进 Z - Score 方法计算其偏移程度，以最终刻画其演化情况，见图 1。

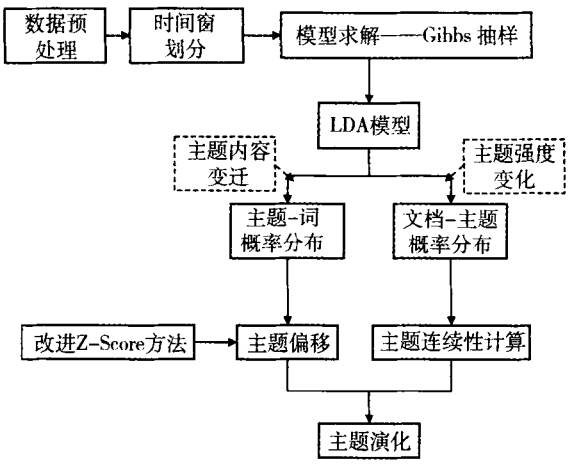

图 1 基于 LDA 模型的主题演化模式

表 1 LDA 模型中使用的符号

符号	描述
D	文档集合
K	主题集合
Nd	文档 d 中词语的个数
$w_{d,i}$	文档 d 中的第 i 个词
$Z_{d,i}$	文档 d 中的第 i 个词的主题
α	LDA 模型的 Dirichlet 先验分布，表示整个文集上主题分布的先验
β	LDA 模型的 Dirichlet 先验分布，表示所有主题上词分布的先验
θ_d	文档 d 上主题的多项式分布
φ_z	主题 z 上词的多项式分布

2.2 主题模型

LDA 模型又称为主题模型，它是一种语言模型，利用层次贝叶斯模型对自然语言进行建模。该模型假设一篇文档由若干潜在主题的概率分布表示，而每一个主题又由一组词的概率分布组成。模型中使用的符号，见表 1。

LDA 模型中的主题由一组统计上相关的词语以及词语在该主题上出现的概率表示。主题 $z = \{(w_1, p(w_1 | z)), \cdots, (w_v, p(w_v | z))\}$，其中 $p(w_v | z)$ 表示主题 z 中词语 w_v 出现的概率。该模型描述了文档的生成过程，其步骤如下[3]：（1）对于文档集合中的文档 $d \in D$，根据文档 d 上主题概率分布的先验：$\theta_d \sim Dir(\alpha)$，得到分布参数 θ_d。（2）对于文档中词的每个主题 $z \in K$，根据主题上词语分布的先验：$\varphi_d \sim Dir(\beta)$，得到分布参数 φ_z。（3）对于文档 d 中的第 i 个词 $w_{d,i}$，跟据概率分布 $z_{d,i} \sim Mult(\theta_d)$，得到主题 $Z_{d,i}$；跟据概率分布 $w_{d,i} \sim Mult(\varphi_z)$，得到词 $w_{d,i}$。LDA 模型，见图 2。

万方数据

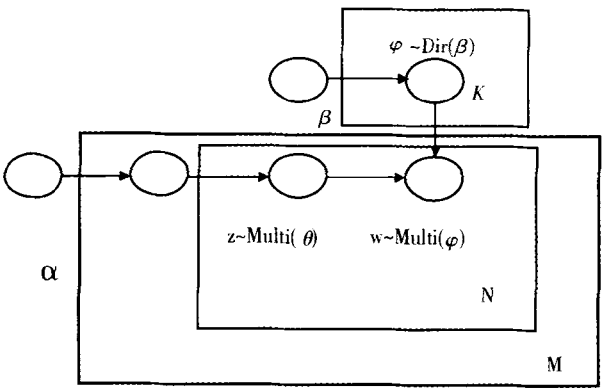

图2 LDA 模型

2.3 模型求解

本文采用 Gibbs 抽样方法求解 LDA 模型中的潜在变量和。Gibbs 抽样是马氏链蒙特卡罗方法（Markov Chain Monte Carlo, MCMC）的一种实现形式，该方法利用每个变量的条件分布，以固定次序从其他变量的条件分布中进行抽样，构造收敛于目标概率分布的马氏链，并从链中抽取接近概率分布值的样本，其过程如下：（1）文档集合中所有词语的数量记为 N，初始化 z_i 为 1 ~ K 之间的某个随机整数，i 从 1 循环到 N，得到马氏链的初始状态。（2）将词语按以下公式分配给主题，i 从 1 循环到 N，得到马氏链的下一个状态。

$$p(z_i = j \mid z_{-i}, w_i) \propto \frac{n^{(w_i)}_{-i,j} + \beta}{n^{\circ}_{-i,j} + V\beta} \cdot \frac{n^{r_i}_{-i,j} + \alpha}{n^{r_i}_{-i,\cdot} + K\alpha}$$

其中，$z_i = j$ 表示将词汇记号 w_i 分配给主题 j，z_{-i} 表示所有 z_N（$N \neq i$）的分配，$n^{(w_i)}_{-i,j}$ 表示词 w_i 被分配给主题 j 的次数，$n^{(\cdot)}_{-i,j}$ 表示分配给主题 j 的所有词语数量，$n^{(r_i)}_{-i,j}$ 表示文档 r_i 中分配给主题 j 的词语数量，$n^{(r_i)}_{-i,\cdot}$ 表示文档 r_i 中所有被分配了主题的词语数量之和。（3）求解 θ 和 φ。当迭代次数充分时，马氏链逼近目标概率分布，获取 z_i 的分布值，得到“特征词 - 主题”矩阵和“文档 - 主题”矩阵。在“特征词 - 主题”矩阵中，行向量表示词汇表中的特征词，列向量表示各个主题，元素值表示某一特征词分配给某一主题的次数；在“文档 - 主题”矩阵中，行向量表示各个文档，列向量表示各个主题，元素值代表文档中的特征词被分配给某一主题的次数。通过以上矩阵计算 θ 和 ϕ 的值：

$$\theta^r_j = \frac{n^{(r_i)}_j + \alpha}{n^{r_i}_{\cdot} + K\alpha}$$

$$\varphi^j_w = \frac{n^{w_i}_j + \beta}{n^{(\cdot)}_j + V\beta} n^r_j$$

表示文档 r 中分配给主题 j 的词语数量，$n^r_{\cdot}$ 为文档 r 中所有被分配了主题的词语数量总和，n^w_j 表示词 w 被分配给主题 j 的次数，$n^{(\cdot)}_j$ 表示分配给主题 j 的所有词语数量。

3 主题间相似性和强度度量

3.1 时间窗划分

本文采用了等距不重叠时间窗进行时间序列划分，然后对划分后的时间序列展开分析：首先确定一个时间间隔，将整个时间序列划分为若干个时间窗，根据文献的出版时间将文献划入相应时间窗。然后应用 LDA 模型对每个时间窗内的文献进行建模。

3.2 主题间相似性和偏移计算

主题间相似性度量。现有研究中衡量两个概率密度的相似度最常用的方法是 KL 距离（Kullback - Leibler Divergence）。KL 距离又称为相对熵（Relative Entropy），衡量相同事件空间里的两个概率分布的差异情况[5]。其公式表示如下：

$$D(P(w \mid s_1) \mid P(w \mid s_2)) = \sum_{wew} P(w \mid s_1) \log \frac{P(w \mid s_1)}{P(w \mid s_2)}$$

主题偏移计算。使用改进 Z - Score 方法计算主题随时间的偏移程度。Z - Score 方法使用平均值和标准差计算集中趋势，而改进 Z - Score 方法使用中值和中值绝对差（Median Absolute Deviation, MAD）计算，更能降低异常值对最终结果的影响。方法如下。（1）计算当前时间窗口与之前各时间窗口中主题间的距离。令 d_{ij} 为当前时间窗口中主题与其他时间窗口中主题间的距离，其中 i 为当前时间窗口中的主题，j 为之前各时间窗口中的主题。采用 Single - pass 方法进行主题间相似度比较计算。常用的 3 种方法有[6]：Single - link, Complete - link, Groupwise - average。这 3 种算法的步骤相似，但是合并两个类的标准不同。Single - link 两个类只要有任意两点距离

万方数据

满足条件即可合并，Complete - link 两个类全部点之间的距离满足条件才能合并，Groupwise - average 则需要两个类所有点之间距离的均值满足条件。在这3种方法中 Groupwise - average 执行效果最好，因为其他两种方法取的是极值，而 Groupwise - average 取的是均值。另外，由于 Single - link 算法只需两个类中有任意两点距离满足阈值即可合并，不需要计算全部点之间的距离，因此运行效率应该是最高的，其次是 Complete - link，最后是 Groupwise - average 算法。因此本文采用 Groupwise - average 算法。（2）改进 Z - Score 公式计算主题间偏移。设 $m = d_{ij}$ 的中值，s 为 MAD，则主题间的偏移量 w 为：$a_{ij} = (d_{ij} - m) / s$。

4 实验结果与分析

4.1 数据预处理

数据源为 SCI 收录 2002 - 2011 年 *Lancet Infectious Diseases* 的全部文献，该刊是 2009 年传染病学领域影响因子最高的期刊。检索式为：so = (" LANCET INFECT DIS")，出版时间为 2002 - 2011 年，共2 848 篇文档。将以上文档以年为单位划分时间窗，将所有文档划分成 10 个时间窗口。

4.2 主题抽取

应用 LDA 模型对每个时间窗口内的文档进行主题抽取，设置主题数目为 10 个，迭代次数设为1 000次进行试验。每个主题使用其概率最大的 10 个词进行表示，以 2002 年的主题为例，见表2。

表2 2002 年主题抽取

主题1	主题2	主题3	主题4	主题5	主题6	主题7	主题8	主题9	主题10
studies	malaria	resistance	infection	hiv	infections	disease	children	disease	tuberculosis
mortality	control	treatment	virus	transmission	patients	factors	vaccine	infectious	diagnosis
therapy	africa	development	immune	infection	associated	host	effective	fever	clinical
evidence	global	drug	viral	available	invasive	pathogenesis	vaccines	influenza	detection
effect	countries	drugs	cause	epidemic	causes	human	developed	diseases	blood
treatment	health	clinical	strategies	prevention	common	bacteria	pneumonia	acute	potential
review	deaths	antibiotics	disease	rates	include	cells	countries	molecular	diagnostic
effects	african	agents	response	increased	antifungal	animal	vaccination	caused	laboratory
outcome	million	antibiotic	hepatitis	haart	review	mechanisms	strains	epidemiological	management
patients	population	including	chronic	potential	major	severe	infection	care	recently

4.3 传染病学领域主题演化分析

根据不同时间窗口内的主题变化情况，分析主题随时间的演化情况。以主题5“HIV”为例，通过表征该主题关键词的概率变化可以直观地反映该主题时间的变化趋势，见表3。

表3 “HIV”关键词变化

时间窗口1（2002 年）	时间窗口2（2003 年）	时间窗口3（2004 年）	时间窗口4（2005 年）	时间窗口5（2006 年）	时间窗口6（2007 年）	时间窗口7（2008 年）	时间窗口8（2009 年）	时间窗口9（2010 年）	时间窗口10（2011 年）
hiv	hiv	hiv	hiv	hiv	hiv	hiv	hiv	hiv	hiv
transmission	development	immune	malaria	data	countries	incidence	therapy	epidemic	infection
infection	effective	transmission	risk	transmission	vaccine	transmission	resistance	strains	diseases
available	transmission	major	data	prevalence	children	prevention	drug	genetic	tuberculosis
epidemic	virus	hcv	prevention	prevention	transmission	population	antiretroviral	mrsa	virus
prevention	viral	viral	pregnancy	strategies	vaccines	women	art	pandemic	update
rates	drugs	response	transmission	hpv	available	prevalence	copies	bcg	cancer
increased	drug	immunity	countries	study	risk	estimates	infected	risky	metaanalysis
haart	increase	research	factors	specific	prevention	infection	testing	sex	pathogenesis
potential	viruses	cells	infection	included	strategies	recent	cryptococcal	forms	usa

万方数据

使用改进 Z - Score 方法计算主题 5 随时间的偏移，刻画其随时间的演化情况，见图 3。

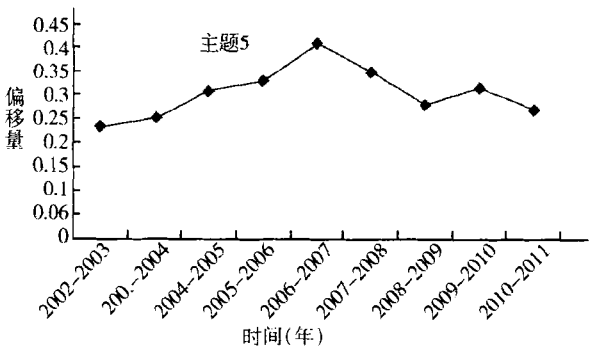

图3 主题 5 “HIV” 随时间变化的演化情况

5 结语

针对文档流开展主题演化分析可以了解信息、事件的发展趋势。本文分析了主题演化的概念及其演化模式，提出了一种基于主题模型识别主题间演化的方法，利用 KL 距离来衡量主题之间的相似度，利用改进的 Z - Score 方法计算主题随时间的偏移程度以刻画其演化情况。并在传染病领域应用 LDA 模型进行主题抽取，开展了主题演化分析。未来将进一步开展主题演化中的“新主题”和“主题消亡”的表征及计量研究，并对中文语料开展基于主题模型的主题演化分析。

参考文献

1 S. Deerwester, S. Dumais, T. Landauer, et al. Indexing by Latent Semantic Analysis [J]. Journal of the American Society of Information Science, 1990, 41 (6): 391 - 407.

2 Thomas Hofmann. Probabilistic Latent Semantic Indexing [C] //Berkeley: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 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 单斌, 李芳. 基于 LDA 模型的话题演化研究方法综述 [J]. 中文信息学报, 2010, 24 (6): 43 - 49.

5 M. Steyvers, T. Griffiths. Probabilistic Topic Models. In: T. Landauer, D. S. McNamara, S. Dennis, et al (Eds.), Handbook of Latent Semantic Analysis [M]. Hillsdale, NJ: Erlbaum, 2007.

6 Vasileios Hatzivassiloglou, Luis Gravano, Ankineedu Maganti. An Investigation of Linguistic Feature and Clustering Algorithms for Topical Document Clustering [C] // Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval, 2000.

关于《医学信息学杂志》启用“科技期刊学术不端文献检测系统”的启事

为了提高编辑部对于学术不端文献的辨别能力，端正学风，维护作者权益，《医学信息学杂志》已正式启用“科技期刊学术不端文献检测系统”，对来稿进行逐篇检查。该系统以《中国学术文献网络出版总库》为全文比对数据库，可检测抄袭与剽窃、伪造、篡改、不当署名、一稿多投等学术不端文献。如查出作者所投稿件存在上述学术不端行为，本刊将立即做退稿处理并予以警告。希望广大作者在论文撰写中保持严谨、谨慎、端正的态度，自觉抵制任何有损学术声誉的行为。

《医学信息学杂志》编辑部

万方数据

基于LDA的主题演化研究

万方数据 WANFANG DATA 文献链接

作者： 李勇， 安新颖， LI Yong， AN Xin-ying
作者单位： 中国医学科学院医学信息研究所 北京100020
刊名： 医学信息学杂志
英文刊名： Journal of Medical Intelligence
年，卷(期)： 2013,34(2)

参考文献(6条)

1. S.Deerwester;S.Dumais;T.Landauer Indexing by Latent Semantic Analysis 1990(06)

2. Thomas Hofmann Probabilistic Latent Semantic Indexing 1999

3. David M.Blei;Andrew Y.Ng;Michael I.Jordan Latent Dirichlet Allocation 2003(03)

4. 单斌;李芳 基于LDA模型的话题演化研究方法综述[期刊论文]-中文信息学报 2010(06)

5. M.Steyvers;T.Griffiths Probabilistic Topic Models 2007

6. Vasileios Hatzivassiloglou;Luis Gravano;Ankineedu Maganti An Investigation of Linguistic Feature and Clustering Algorithms for Topical Document Clustering 2000

引用本文格式：李勇.安新颖.LI Yong.AN Xin-ying 基于LDA的主题演化研究[期刊论文]-医学信息学杂志 2013(2)