一种基于LDA的在线主题演化挖掘模型

崔 凯 周 斌 贾 焰 梁 政
(国防科学技术大学计算机学院 长沙 410073)

摘 要 基于文本内容的隐含语义分析建立在线主题演化计算模型,通过追踪不同时间片内主题的变化趋势进行主题演化分析。将Latent Dirichlet Allocation(LDA)模型扩展到在线文本流,建立并实现了在线LDA模型;利用前一时间片的后验概率影响当前时间片的先验概率来维持主题间的连续性;根据改进的增量Gibbs算法进行推理,获取主题-词和文档-主题的概率分布,利用Kullback Leibler(KL)相对熵来衡量主题之间的相似度,从而发现主题演化中的"主题遗传"和"主题变异"。实验结果表明,该模型能从互联网语料中找出主题的演化趋势,具有良好的效果。

关键词 主题模型,LDA,演化,舆情
中图法分类号 TP310 **文献标识码** A

LDA-based Model for Online Topic Evolution Mining

CUI Kai ZHOU Bin JIA Yan LIANG Zheng
(School of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract A computational model for online topic evolution mining was established through a latent semantic analysis process on textual data. Topical evolutionary analysis was achieved by tracking the topic trends in different time-slices. In this paper, Latent Dirichlet Allocation (LDA) was extended to the context of online text streams, and an online LDA model was proposed and implemented as well. The main idea is to use the posterior of topic-word distribution of each time-slice to influence the inference of the next time-slice, which also maintains the relevance between the topics. The topic-word and document-topic distributions are inferenced by incremental Gibbs algorithm. Kullback Leibler (KL) relative entropy is uesd to measure the similarity between topics in order to identify topic genetic and topic mutation. Experiments show that the proposed model can discover meaningful topical evolution trends both on English and Chinese corpus.

Keywords Topic model, LDA, Evolution, Public opinion

1 引言

随着互联网的日益普及和网络公民意识的觉醒,网络不仅成为民众获取和发布信息的主要渠道,而且成为参政议政和政府监督的重要平台。网上舆论在"汶川地震"、"邓玉娇事件"、"天价烟事件"等中显示了巨大的影响力,受到各级决策者的关注。舆情分析的关键是通过对互联网上的海量文本数据进行分析,掌握主题(Topic,也有部分研究者译为话题)的演化趋势,做出及时正确的预测,供决策者参考。

目前国外在该领域的研究主要是DARPA在1996～1997年的专项资助及随后的美国国家标准技术研究所(NIST)资助并主持的一系列话题检测与追踪系列评测会议[1,2];IBM Almaden研究中心基于博客数据进行了大量实证研究,发现了话题的若干组成特征,区分了闲谈式(chatter)和突发式(spike)两类不同性质的话题,给出了两类话题的结构特征描述以及基于统计显著性差异的话题判定算法[3];Zhou等[4]将话题演化视为Markov过程,通过Markov转移矩阵为话题演化建模,关注话题演化在多大程度上受其后台隐含用户实体间交互的影响;伊利诺伊大学的翟成祥定义了内容演化和强度演化两个主题演化模式[5]和相关文本挖掘的混合模型[6]。

国内关于网络话题演变机制方面的研究也正起步。哈尔滨工业大学、中科院计算所、国防科技大学等单位都进行了相关研究。于满泉等研究了层次化话题识别技术,将自然语言处理与信息检索技术相结合,提出了单粒度话题识别方法和基于多层聚类的MLCS算法,对话题进行层次化组织[7]。王永恒等在面向汉语短文的话题识别系统研究方面展开了相关研究工作[8]。张立等[9]通过对我国某网络论坛数据进行分析处理,发现该网络的度分布为幂律分布,有明显的无标度特征。胡勇等对意见领袖的形成模型做了定量研究,提出了意见领袖的属性矩阵[10]。

统计主题模型(Statistic Topic Model)近年来得到了充分

到稿日期:2009-12-28 返修日期:2010-03-16 本文受国家自然科学基金重点项目(60933005),面上项目(60873204)资助。
崔 凯 男,硕士,主要研究方向为数据挖掘、机器学习,E-mail:cuikai186@gmail.com;周 斌 男,副研究员,硕士生导师,主要研究方向为数据挖掘、信息检索、分布计算;贾 焰 女,教授,博士生导师,主要研究方向为数据库、网络安全、分布计算;梁 政 男,博士,主要研究方向为数据挖掘。

的重视和深入的研究。普林斯顿大学的 David M. Blei 首先提出了 LDA 模型[11]，用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文档的主题混合比例，模拟文档的产生过程；之后考虑主题之间的相关性提出了 CTM 模型[12]、考虑时间信息提出了动态主题模型[13,14]。LDA 的扩展模型还包括作者-主题模型[15]、作者-角色-主题模型[16]、OLDA 模型[17]等。石晶等研究了基于 PLSA 和 LDA 的文本分割算法[18,19]。俞辉等研究了基于 PLSA 的多文档自动文摘和用户聚类算法[20,21]。上述相关研究中 OLDA 模型与本文的研究最相似，但其主要是基于英文文本，对权重的设置过于复杂。本文建立在线 LDA 模型来对文本进行主题演化分析，利用中文分词功能[22]，实现了对中文和英文两种文档的分析处理。

本文第 1 节介绍相关研究背景；第 2 节提出主题演化计算模型及增量 Gibbs 算法；第 3 节阐述主题强度度量和相似性度量；第 4 节进行实验结果分析；最后总结全文。

2 主题演化挖掘模型

2.1 主题演化模型

演化(Evolution)又称进化，指生物在不同世代之间具有差异的现象以及解释这些现象的各种理论。演化的主要机制是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。自然选择的过程会使物种的特征被保留或是淘汰，甚至使新物种诞生或原有物种消失。

主题演化同样具有“主题遗传”和“主题变异”特性。“主题遗传”是指主题具有一定的稳定度，“主题变异”是指主题具有一定的差异性。本文采用主题-词概率密度来表示一个主题(见图 1)。“主题遗传”和“主题变异”就转变成主题-词概率分布随着时间的变化具有一定的稳定性和差异性，用主题特征关键词概率的变化来进行表示。生物学中的遗传和变异取决于生物成长环境的变化和不同物种父对子的遗传程度。“主题遗传”和“主题变异”取决于语料的变化和前一主题对当前同一主题的影响程度。本文定义一个 W 权重来表示主题的“遗传度”，这个参数值取决于具体的应用领域。比如在新闻语料中，话题的变化通常比较频繁，W 值较小；科学研究中的一般话题具有很大的稳定性，W 值较大。W 由用户自定义，其值为非负数，W 为 0 表示话题之间无遗传，W 越大表示话题越稳定，传播过程中语料的影响就越小。

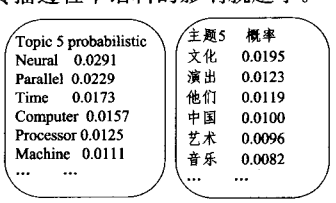


图 1 主题-词概率密度表示

本文将 LDA 模型扩展到在线文档，建立在线 LDA 模型。其输入是一系列随时间变化的文档块，输出是对应文档块的主题-词概率分布，用 ϕ 表示；文档-主题概率分布用 θ 表示。本文用主题在不同时间块所占的比重即 $\bar{\theta}$ 来表示“主题遗传”；用 KL 相对熵检测主题-词概率分布的差异来表示“主题变异”。

主题演化分析的主要任务就是从一系列文本流中抽取出隐含的主题并发现其演化规律。首先采用滑动窗口技术，把文本流划分成一个个时间片，各时间片内的文档数目可以不同，不同时间片内的文档也可以重复。然后对每个时间片内的文档采用 LDA 模型进行建模，抽取出 K 个主题，得出主题-词和文档-主题的概率分布。并把前一个时间片的主题-词概率分布加上适当的权重 W 作为当前时间片的先验概率，然后根据主题-词和文档-主题概率分布随时间的变化得出主题演化图和主题强度变化图两种基本的主题演化模式(见图 2)。

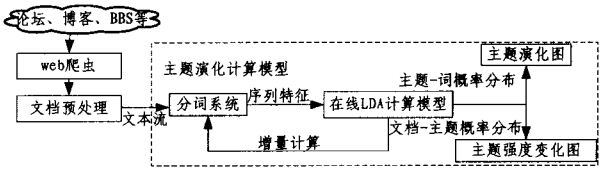


图 2 话题演化系统架构

进行文本挖掘的第一步是对文本进行分词，使之形成序列特征文档。英文分词按空格对单词进行划分，然后去除停用词，这里采用正则表达式来实现。中文分词比较复杂，中文中的词相当于英文中的词组，语义层次更深一层，采用开源的基于 lucene 的 imdict-chinese-analyzer 中文分词模块。该分词器算法基于隐马尔科夫模型，分词速度平均为 259517 汉字/s[22]。

2.2 在线 LDA 计算模型

LDA 是一个三层贝叶斯概率模型，包含词、主题和文档三层结构。LDA 将每个文档表示为一个主题混合，每个主题是固定词表上的一个多项式分布(Multinomial Distribution)。LDA 假设文档由一个主题混合产生，同时每个主题是在固定词表上的一个多项式分布；这些主题被集合中的所有文档共享；每个文档有一个特定的主题混合比例(Topic Proportion)，其从 Dirichlet 分布中抽样产生。作为一种产生式文档模型，用 LDA 提取文档的隐含语义结构和文档表征已经成功地应用到很多文本相关的领域。

LDA 模型中直接计算参数 ϕ 和 θ 是不可能的，一般都是通过最大似然估计的方法来进行参数估计。经典的算法有 Mean Field Variational 算法、Expectation Propagation 算法、Variational Inference 算法和 Gibbs 算法，其中 Gibbs 算法由于其实现简单、计算速度快而得到了广泛的应用。

面向在线文本流处理时，需对 LDA 模型进行扩展，建立在线 LDA 计算模型，在增量计算的同时保持主题间连续性并检测主题间差异性。首先定义时间 t 内的互联网文档集合 $S_t=\{d_1,d_2,\cdots,d_{Mt}\}$，可变尺寸 M_t。滑动窗口的尺寸依赖于用户的需求和具体的应用领域以及语料分析的粗粒度，一般以天、月、年等时间顺序进行划分。同一时间块内的文档是可交换的，不同时间块内的文档是不可交换的。本文采用后验和先验概率来保持主题间的连续，即用前一时间块计算出的主题-词的后验概率 ϕ_{t-1} 乘上权重 W 作为当前时间片的先验概率 ϕ_t，即 $\phi_t=\phi_{t-1}W$，建立在线 LDA 计算模型(见图 3)。

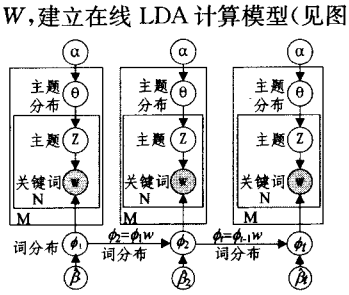


图 3 在线 LDA 文档模型

在线 LDA 计算模型可以从文本流中抽取主题、识别突发主题和主题随时间变化的趋势，并且能够随文本流增量构造和实时更新(主题-词概率密度和文档-主题混合)。处理当前时间片时不需要获得该时间片前的数据，节约了内存，使之能够处理大规模语料，适合在线分析的环境。文本流中会持续不断地引入新的词汇，本文假定新词汇在以前所有时间片的文档集合中出现的次数为 0。在线 LDA 计算模型中文档的生成过程(见表 1)，其中使用的 W 根据用户设定的 w 变化而来，其计算公式为：

$$W=(Token)_t w/(Token)_{t-1} \tag{1}$$

式中，$Token$ 表示文档集中的词数。对用户设定的 w 进行转换的目的是避免时间片内词数的突变带来效果的突变，使该模型对语料环境的变化具有更广的适应度，确切地反映用户设定的 w 值表征的历史对当前文档集的影响力。

模型中要推理的参数是主题-词概率分布 ϕ 和文档-主题概率分布 θ。这里对 Gibbs 算法进行改进，使之适合增量计算模式，以便对本模型进行参数估计。

表 1　在线 LDA 文本生成过程

在线 LDA 文本生成过程
1. For each topic $k=1,2,\cdots,k$
2. If $k=1$
3. Draw $\phi_1 \sim Dir(\cdot\|\beta_1)$
4. else
5. Computer $\phi_t=\phi_{t-1}w$
6. For each document d
a) Draw $\theta_d \sim Dir(\cdot\|\alpha^t)$
b) For each word token ω_i in document d
i. Draw z_i from multinomial θ_d; $p(z_i\|\alpha^t)$
ii. Draw ω_i from multinomial φt: $p(\omega_i\|z_i,\beta_{z_i})$

2.3　增量 Gibbs 算法

在 LDA 模型中，为了获取词汇的概率分布，Gibbs 算法没有将 ϕ 和 θ 作为参数直接计算，而是考虑词汇对于主题的后验概率 $P(z|w)$，利用抽样间接求得 ϕ 和 θ 的值。Markov chain Monte Carlo(MCMC)系一套从复杂的概率分布抽取样本值的近似迭代方法。Gibbs 抽样作为 MCMC 的一种简单实现形式，其目的是构造收敛于某目标概率分布的 Markov 链，并从链中抽取被认为接近该概率分布值的样本[19]。本文中的在线 LDA 计算模型需要把前一时间片中求得的 ϕ_{t-1} 加上适当的权重 W 作为当前时间片样本抽样的先验概率，这样时间片 t 时刻的后验概率为 $P_t(z_i=j|z_{-i},w_i)$，即目标函数计算公式如下：

$$P_t(z_i=j|z_{-i},w_i)=\frac{(n_{-i,j}^{(w_i)})_t+w(n_{-i,j}^{(w_i)})_{t-1}+\beta}{(n_{-i,j}^{(\cdot)})_t+w(n_{-i,j}^{(\cdot)})_{t-1}+V\beta}\cdot\frac{(n_{-i,j}^{(d_i)})_t+\alpha}{(n_{-i,\cdot}^{(d_i)})_t+T\alpha}\Big/\sum_{j=1}^{T}\frac{(n_{-i,j}^{(w_i)})_t+w(n_{-i,j}^{(w_i)})_{t-1}+\beta}{(n_{-i,j}^{(\cdot)})_t+w(n_{-i,j}^{(\cdot)})_{t-1}+V\beta}\cdot\frac{(n_{-i,j}^{(d_i)})_t+\alpha}{(n_{-i,\cdot}^{(d_i)})_t+T\alpha} \tag{2}$$

式中，$w(n_{-i,j}^{(w_i)})_{t-1}$ 是上一时间片内分配给主题 j 与 w_i 相同的词汇个数，$w(n_{-i,j}^{(\cdot)})_{t-1}$ 是分配给主题 j 的所有词汇个数；W 为权重，值为非负 $z_i=j$ 表示将词汇记号 w_i 分配给主题 j，这里 w_i 被称为词汇记号是因为其不仅代表词汇 w，而且表示的是在文本中的位置有关的词汇；，z_{-i} 表示所有 $z_k(k\neq i)$ 的分配。$n_{-i,j}^{(w_i)}$ 是分配给主题 j 与 w_i 相同的词汇个数；$n_{-i,j}^{(\cdot)}$ 是分配给主题 j 的所有词汇个数；$n_{-i,j}^{(d_i)}$ 是 d_i 中分配给主题 j 的词汇个数；$n_{-i,\cdot}^{(d_i)}$ 是 d_i 中所有被分配了主题的词汇个数；所有的词汇个数均去掉这次 $z_i=j$ 的分配。

增量 Gibbs 抽样算法详述如下[25]：

(1)z_i 被初始化为 1 到 T 之间的某个随机整数。i 从 1 循环到 N，N 是语料库中所有出现于文本中的词汇记号个数。此为马尔科夫链的初始状态。

(2)i 从 1 循环到 N，根据式(2)将词汇分配给主题，获取马尔科夫链的下一个状态。

(3)迭代第(2)步足够次数以后，认为马尔科夫链接近目标函数分布，遂取 z_i(i 从 1 循环到 N)的当前值作为样本记录下来。

舍弃词汇记号，以 w 表示唯一性词，对于每一个样本即时间片内的文档集合，可以按式(3)估算 ϕ 和 θ 的值，其中 $n_j^{(w)}$ 表示词汇 w 被分配给主题 j 的频数；$n_j^{(\cdot)}$ 表示分配给主题 j 的所有词数；$n_j^{(d)}$ 表示文本 d 中分配给主题 j 的词数；$n_\cdot^{(d)}$ 表示文本 d 所有被分配了主题的词数。

$$\begin{aligned}(\hat{\varphi}_w^{(z=j)})_t&=\frac{(n_j^{(w)})_t+w(n_j^{(w)})_{t-1}+\beta}{(n_j^{(\cdot)})_t+w(n_j^{(\cdot)})_{t-1}+V\beta}\\(\hat{\theta}_{z=j}^{(d)})&=\frac{(n_j^{(d)})_t+\alpha}{(n_\cdot^{(d)})_t+T\alpha}\end{aligned} \tag{3}$$

3　主题间相似性和强度度量

根据在线 LDA 模型可以获得一系列主题-词概率分布和文档-主题概率分布，依据这两个概率分布就可以检测到主题传播过程中的“主题遗传”和“主题变异”。本文采用 KL 相似性距离度量标准来刻画主题-词概率分布之间的相似度、检测主题变化的差异，然后根据差异的情况来发现突发主题，找出主题变化趋势。对主题的稳定性用主题在时间片中所占的比重，即 θ 的平均值来代表主题的强度，绘制主题强度变化图。

3.1　主题间相似性度量

衡量两个概率密度的相似度度量标准最常用的就是 KL 距离。KL 距离也叫相对熵、交叉熵，是对样本概率为 P 的信息使用概率编码后期望的额外增加的比特数，经常用来衡量两个概率密度之间的差距。其表示如下：

$$D(P(w|s_1)||P(w|s_2))=\sum_{w\in W}P(w|s_1)\log\frac{P(w|s_1)}{P(w|s_2)} \tag{4}$$

标准 KL 距离是非对称的，值是非负的。当值为 0 时，表示两个概率密度完全相同。根据文献[23]KL 有 4 种变形，其计算方式如下：

1)标准 KL 距离：$KL(A,B)=D(A||B)$

2)对称 KL 距离：$KL(A,B)=D(A||B)+D(B||A)$

3)clarity-adjusted：$KL_{c,1}(A,B)=D(A||B)-\text{Clarity}(A)$

4)对称 clarity-adjusted ：$KL_{c,2}(A,B)=KL_{c,1}(A,B)+KL_{c,1}(B,A)$

本文采用标准 KL 距离和对称 KL 距离两种形式来衡量主题之间的相似度。通过实验比较，其刻画主题的差异性区别不大，采用哪一种距离度量标准都可以。需要注意的是在计算过程中，文本流不断地引入新的词汇，前述假设当前时间块中的新词在以前的所有时间片中的出现次数为 0，并在处理不同时间片的过程中维持一个一致且增量更新的词典，使不同时间片的主题-词概率分布在统一的概率空间上，方便进行 KL 距离的计算和比较。

3.2 主题强度度量

衡量话题稳定程度最直接的标准就是观察主题强度随时间不断变化的趋势。Gibbs 抽样算法依据式(3)获得的$\hat{\theta}$表示每个文本中主题的混合程度，把文本扩展到每个时间片的文档集，同样可以求出该时间片的主题混合程度$\bar{\theta}_t$，表示当前时间片内语料的主题混合平均程度。用每个主题的$\bar{\theta}_k$表示当前时间片中某主题的强度[25]，可以得出文本流中一系列主题强度的不同值。依据此可绘制出主题强度变化曲线，从宏观上基于更深层次的语义分析发现主题的变化趋势。

在线 LDA 模型把每篇文章看成多个主题的混合，即每篇文章都可以按概率属于多个主题，类似于不确定性聚类。求出的主题$\bar{\theta}$代表了当前整个语料中该主题的混合比例，是建立在词的区分基础上的，有着更深层次的语义分析，更加客观和准确。

4 实验结果及讨论

本文实验使用英文数据集 NIPS[24]和从天涯杂谈板块抓取的中文数据集进行分析。

NIPS：该数据集共包括 1988－2001 年共 14 年会议 1958 篇文章的全文数据，大小为 35.1M，文件格式为文本格式，以年为单位划分成 14 个文件夹。

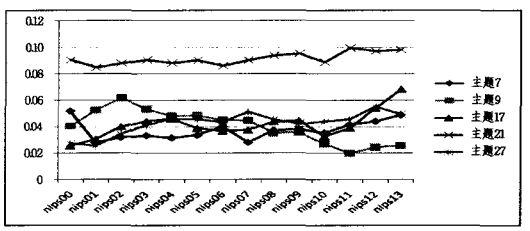
中文数据集：天涯社区中天涯杂谈板块 2008 年 10，11 两个月的数据，原始文件 1.3G，经预处理后文件大小 48M，18310 个文件，以周为单位划分成 8 个时间片。

实验环境：机器 CPU：AMD 双核 3800，内存 2G，硬盘 250G。

4.1 nips 数据集实验分析

设置主题数目为 30 个主题、迭代次数为 500 次、权重 W 为 0.3 进行实验。

挑选出主题 7，9，17，21，27 进行展示(见图 5)，每个主题以其概率最大的前 15 个词进行标识。因为主题的概率分布是不断变化的，以最后一个时间片 nips13 即 2001 年的进行展示。通过图表可以看出主题强度的变化趋势，其中主题 21 以较高的强度区别于其他主题。通过代表主题的关键词概率可以看出这些词汇具有一定的普遍性，即背景噪声词汇。主题 17，27 呈上升趋势，主题 9 呈下降趋势，主题 7 波动不大。



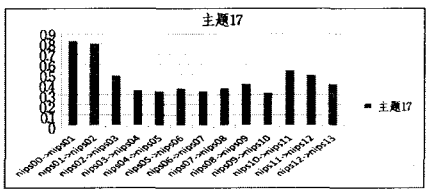
Topic 7	Topic 9	Topic 17	Topic 21	Topic 27
bound	units	training	results	learning
theorem	learning	kernel	figure	distribution
function	input	data	values	probability
functions	hidden	margin	large	error
bounds	weights	set	number	random
case	layer	support	show	sample
proof	training	vector	small	algorithm
result	network	error	shows	examples
xi	unit	svm	obtained	distributions
positive	outpu	test	standard	bayeslan
linear	weight	algorithm	note	log
convergence	test	feature	similar	prior
approximation	trained	classification	university	generalization
lemma	models	classifier	function	statistical
theory	error	classifiers	present	xi

图 5 主题强度变化图

对主题 17 的演化趋势进行分析(见图 7)，其中柱形图表示的是主题 17 在相邻时间块内的 KL 距离，柱形图中的长度表示 KL 距离，即主题 17 随时间的变化程度。通过观察柱形图中的突变可以发现主题 17 的变化，例如在 nips10->nips11 中距离突然增大，表明主题 17 在这段时间内发生了重大变化。通过对代表该主题的关键词概率对比可以发现(见图 6)，属于 SVM 的相关关键词的概率明显增大，并在 nips12 和 nips13 中继续呈上升趋势，这说明主题 17 的内容逐渐演化成 SVM。通过代表每个时间块主题 17 的最大概率的文章标题可以更直观地感觉到其关注侧重点的变化趋势。

Topic 17	Topic 17	Topic 17	Topic 17	Topic 17	Topic 17	Topic 17
training	training	training	training	training	training	training
set	set	set	set	set	set	kernel
performance	test	performance	test	data	kernel	data
network	performance	test	data	test	data	margin
test	trained	data	error	error	support	set
trained	examples	error	performance	kernel	test	support
data	network	generalization	examples	examples	svm	vector
generalizat	data	trained	trained	margin	error	error
net	generalization	examples	generalization	performance	vector	svm
architecture	ensemble	ensemble	classifiers	support	margin	test
character	character	network	size	sets	examples	algorithm
examples	sets	size	classifier	classifier	machines	feature
errors	size	classifiers	sets	generalization	sets	classification
input	characters	sets	ensemble	vector	algorithm	classifier
machine	input	obs	machine	classification	classification	classifiers
predictions	based	feature	input	svm	performance	kernels
size	error	validation	network	adaboost	regression	examples
hand	errors	number	number	classifiers	adaboost	space

图 6 主题 17(nips07-nips13)关键词变化图



nips00 A "Neural" Network that Learns to Play Backgammon
nips01 Connectionist Learning of Expert Preferences by Comparison Training
nips02 Optimal Brain Damage
nips03 A B P ANN Commodity Trader
nips04 Recognizing Overlapping Hand Printed Characters by Centered Object Integrated Segmentation
nips05 Training a Net to Make Saccades and Recognize Handwritten Characters
nips06 Signature Verification using a "Siamese" Time Delay Neural Network
nips07 Limits on Learning Machine Accuracy Imposed by Data Quality
nips08 Human Reading and the Curse of Dimensionality
nips09 Combined Weak Classifiers
nips10 Training Methods for Adaptive Boosting of Neural Networks
nips11 Using Analytic QP and Sparseness to Speed Training of Support Vector Machines
nips12 Leveraged Vector Machines
nips13 Fast Training of Support Vector Classifiers

图 7 主题 17 演化图

4.2 天涯杂谈数据集

设置主题数目为 30 个主题、迭代次数为 500 次、权重 W 分别为 0.5 进行实验。

对于主题 13(见图 8)，可以看出主要是食品产品质量安全类别的文章，在时间片 1 和 2 内焦点的是奶粉中毒事件。而在时间片 3 开始，奶粉中毒事件关注度相对减弱，被微软黑屏事件所取代，成为关注的焦点。从其关键词的变化和该主题概率最大的两篇文章可以很明显地看出该主题热点的趋势和变化。

时间片1	时间片2	时间片3	时间片4	时间片5	时间片6	时间片7	时间片8
企业	企业	企业	产品	企业	产品	产品	产品
奶粉	奶粉	产品	企业	产品	产品	企业	企业
产品	产品	盗版	问题	生产	问题	品牌	问题
生产	生产	生产	盗版	问题	品牌	问题	生产
品牌	问题	奶粉	生产	安全	生产	生产	品牌
问题	质量	问题	质量	质量	公司	消费者	消费者
食品	品牌	软件	软件	消费者	安全	安全	技术
质量	食品	用户	奶粉	事件	销售	销售	公司
安全	牛奶	质量	安全	销售	质量	使用	使用
牛奶	安全	安全	消费者	奶粉	消费者	公司	销售
事件	事件	使用	使用	使用	奶粉	质量	质量
公司	市场	系统	销售	市场	使用	技术	安全
市场	标准	牛奶	用户	食品	市场	市场	市场
配方	公司	品牌	市场	公司	事件	事件	系统
标准	消费者	公司	公司	品牌	技术	行业	软件
消费者	销售	食品	牛奶	技术	行业	奶粉	奶粉

时间片1 高度赞扬广东省牛奶公司.txt
三元能否为自己喊声冤？.txt
时间片2 卫生部首次制定三聚氰胺限量 婴儿乳粉为1mg（转）.txt
三聚氰胺可以加进奶粉？！.txt
时间片3 10月21日，电脑公司，排队装系统…….txt
注意！！！微软的Windows正版增值计划的全面说明解惑(持续更新).tx
时间片4 支持微软，自觉黑屏！（图）.txt
淘宝混了2年，揭发一些黑幕，今天先揭发当年红遍大江南北的减肥细带
饲料加三聚氰胺是公开秘密 水产业可能是重灾区.txt
时间片5 三鹿被三元收购，取个三牛，三马，三驴继续生产奶粉吧.txt
时间片6 急盼请教一下,请问这个密保有没有市场前途.txt
硬盘应该可以恢复数据的阿！！！！.txt
时间片7 10月全国各地鸡价～太贵了.txt
《延生护宝酒考证洋酒的酿造工艺是中国发明的》.txt
时间片8 职业打假！飞科牌电剪斗和电吹风冒称中国驰名商标.txt
中国的洲际导弹.txt

图 8 主题 13 主题演化图

(下转第 193 页)

[3] 沈海波，洪帆. 访问控制模型研究综述[J]. 计算机应用研究，2005(6)：9-11

[4] Jaehong P, Ravi S. The UCONABC Usage Control Model [J]. ACM Transactions on Information and System Security (TISSEC), 2004(10): 1-47

[5] Zhang Xin-wen, Parisi-Presicce, Ravi S. Formal model and policy specification of usage control [J]. ACM Transactions on information and System Security, 2005, 8(4): 351-387

[6] Osborn S, Sandhu R, Munawer Q. Configuring Role-based Access Control to Enforce Mandatory and Discretionary Access Control Policies [J]. ACM Transactions on Information and Systems Security(TISSEC), 2000, 3(2): 85-106

[7] Pretschner A, Hilty M, Basin D. Distributed usage control [J]. Communications of the ACM, 2006, 49(9): 39-44

[8] 黄云，周敏. 从电子病历的发展历程谈医疗安全管理中的风险管理[J]. 中国医院管理，2007，27(10)

[9] 林彩霞. 电子病历档案首页存在的问题及改进意见[R]. 广东档案业务探讨，2009

(上接第 159 页)

对于主题 11(见图 9)，可以看出主要关注的是阎崇年被打事件的相关讨论，并在 8 个时间片内都保持了相当高的热度，而且从历史文化相关的主题中也有一定量的阎崇年被打的帖子。可以分析出关于此事件的讨论有往历史文化方面扩展的趋势，发现该主题讨论的不同侧面。根据主题 11 的主题在 8 个时间片内的差异变化柱形图，可以看出其十分稳定，在 2008 年 11 月到 12 月之间是个热点话题。

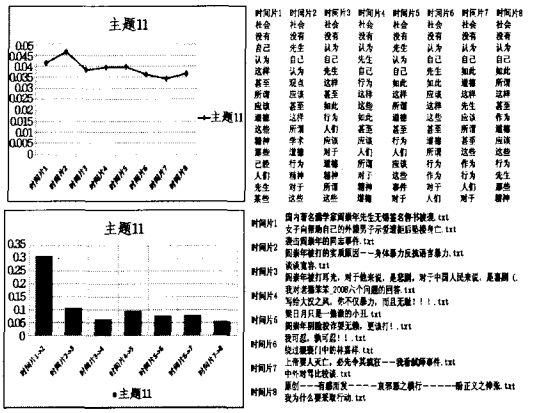


图 9 主题 11 主题演化趋势图

结束语 本文首先分析并定义了主题演化的概念和蕴含在其中的演化模式，然后建立并实现了在线 LDA 模型，最后通过实验表明此模型在主题演化分析上有着良好的效果，能够发现主题传播过程中的“主题遗传”和“主题变异”现象，基于更深层次的文本语义分析的概率主题模型具有坚实的理论基础，更加客观和准确。但本文只考虑了文本的内容和时间属性，没有考虑文本的作者、出处、链接、回帖等结构属性信息。如何把这些属性信息融合进去，改进模型的效果；如何采用基于云的 Gibbs 算法改进模型的效率，是我们下一步要努力的方向。

参 考 文 献

[1] CaliforniaAllan J, Carbonell J, Doddington G, et al. Topic detection and tracking pilot study: Final report[C]// Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. February 1998

[2] The 2002 Topic Detection and Tracking (TDT2002) Task Definition and Evaluation Plan[EB/OL]. ftp://jaguar. ncsl. nist. gov//tdt/tdt2002/evalplans/TDT02. Eval. Plan. v1. 1. pdf

[3] Gruhl D, Guha R V, Liben-Nowell D, et al. Information diffusion through blogspace[C]// WWW. 2004: 491-501

[4] Zhou Ding, Ji Xiang, Zha Hongyuan, et al. Topic evolution and social interactions: how authors effect research[C]// CIKM. 2006: 248-257

[5] Mei Qiaozhu, Zhai Chengxiang. Discovering Evolutionary Theme Patterns from Text-An Exploration of Temporal Text Mining [C]// Proceedings of KDD'05. 2005

[6] Mei Qiaozhu, Zhai Chengxiang. A Mixture Model for Contextual Text Mining[C]// Proceedings of KDD'06. 2006

[7] 于满泉，骆卫华，许洪波，等. 话题识别与跟踪中的层次化话题识别技术研究[J]. 计算机研究与发展，2006，43(3)：489-495

[8] 王永恒，贾焰，杨树强. 面向汉语短文的话题识别系统研究[C]// 第 21 届全国数据库年会(NDBC2004). 计算机科学，31(10 增刊)

[9] 张立，刘云. 网络舆论传播的无标度特性及其衰减模型的研究[J]. 北京交通大学学报，2008(4)

[10] 胡勇，张羽斌，王祯学，等. 网络舆论形成过程中意见领袖形成模型研究[J]. 四川大学学报：自然科学版，2008(4)

[11] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022

[12] Blei D, Lafferty J. A correlated topic model of science[J]. Annals of Applied Statistics. 1: 1 17-35

[13] Blei D, Lafferty J. Dynamic topic models[C]// Proceedings of the 23rd International Conference on Machine Learning. 2006

[14] Wang C, Blei D, Heckerman D. Continuous time dynamic topic models[C]// Uncertainty in Artificial Intelligence(UAI). 2008

[15] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]// Proceedings of the 20th International Conference on Uncertainty in AI. July 2000

[16] McCallum A, Wang A, Corrada-Emmanuel. Topic and role discovery in social networks[C]// Proceedings of 19th Joint Conference on Artificial Intelligence. 2005

[17] Alsumaitm L, Barbara D, Domeniconi C. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking[C]// IEEE International Conference on Data Mining. 2008: 3-12

[18] 石晶，胡明，石鑫，等. 基于 PLSA 模型的文本分割[J]. 计算机研究与发展，2007(2)

[19] 石晶，胡明，石鑫，等. 基于 LDA 模型的文本分割[J]. 计算机学报，2008(10)

[20] 俞辉. 基于 LSA 和 pLSA 的多文档自动文摘[J]. 计算机工程与科学，2009，31(9)

[21] 俞辉. 基于 PLSA 模型的 Web 用户聚类算法研究[J]. 计算机工程与科学，2008，30(7)

[22] imdict-chinese-analyzer[EB/OL]. http://code. google. com/p/imdict-chinese-analyzer/

[23] Lavrenko V, Allan J, DeGuzman E, et al. Relevance models for topic detection and tracking[C]// Proceedings of the Second International Conference on Human Language Technology Research. March 2002: 24-27

[24] NIPS dataset[EB/OL]. http://nips. djvuzone. org/txt. html

[25] Griffiths T L, Steyvers M. Finding scientific topics[C]// Proceeding of the National Academy of Sciences. 2004: 5228-5235