一种面向涌现的比较性话题模型*

谭文堂,王桢文,殷风景,葛 斌,肖卫东
(国防科技大学 信息系统工程重点实验室,湖南 长沙 410073)

摘 要:提出一种CDCMLDA生成模型来实现跨文本集的话题分析,采用狄利克雷组合多项式模型(Dirichlet Compound Multinomial ,DCM)对文本集中词的涌现现象进行建模,把DCM模型和LDA结合起来分析文本集之间话题的差异,采用蒙特卡罗期望最大化方法进行参数推导。在多个实际数据集中通过定性和定量的方法对模型进行评价,实验表明,模型不仅能够发现不同文本集间的异同,而且在模型困惑度指标上相对当前两种主要跨文本集的话题模型具有明显的优势。

关键词:比较性文本挖掘;涌现;话题模型;CDCMLDA模型
中图分类号: **文献标志码**:A **文章编号**:1001-2486(2013)04-0146-10

A comparative topic model for words burstiness

TAN Wentang, WANG Zhenwen, YIN Fengjing, GE Bin, XIAO Weidong
(Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China)

Abstract: State-of-the-art cross collections topic models suffer from the serious flaw that it cannot capture the tendency of words to appear in bursts. Based on LDA (Latent Dirichlet Allocation), a topic model CDCMLDA (Cross-collection Dirichlet compound multinomial Latent Dirichlet Allocation), which models the burstiness phenomena of words using Dirichlet compound multinomial (DCM) distribution, was proposed. A Monte Carlo Expectation Maximization algorithm for model inference was presented. A variety of qualitative and quantitative evaluations of CDCMLDA were performed, which shows that CDCMLDA not only discovers the common and unique aspects on topics, but also improves the model perplexity compared with the two cross-collection topic models.

Key words: comparative text mining; burstiness; topic model; CDCMLDA model

1 研究背景

比较性文本挖掘旨在发现可比或相似文本集之间语义结构之间的差异,如话题在不同时间、地域、文化的人群中所表现出来的差异。所谓可比文本集是指讨论类似话题的多个文本集。在互联网飞速发展的今天,比较性文本挖掘具有十分重要的现实意义。科研人员可通过它分析某个领域的研究热点在几年之间的变化趋势;决策者则需要了解在有关措施实施之前与之后民众态度的变化;企业通过分析相关用户的博客可了解不同地区的人对于同一个产品的评价的不同,不同年龄阶段的人消费观念的差异等。

当前比较性文本挖掘的模型主要有:CCMix(Cross-Collection Mixture)模型和CCLDA(Cross-Collection LDA)模型,两个模型在一定程度上解决了跨文本集的比较性文本挖掘问题,但由于没有考虑词的涌现特征,模型描述能力和预测能力较弱[1-4]。本文根据词的涌现规律[5],提出一种跨文本集的CDCMLDA模型,采用狄利克雷合成多项式分布(Dirichlet Compound Multinomial, DCM)来对在不同文档集之间词的涌现现象进行建模,基于此来分析多个文本集之间的差异。

2 相关工作

文本处理一般通过对文本建立向量空间模型进行文本分析、索引和检索,但向量空间模型没有考虑同义词、近义词以及一词多义等语义特征,因此我们需要对文本的语义进行更深层次的分析。潜在语义索引(Latent Semantic Index, LSI)是这方面的先驱[6]。LSI把文档集表示为"词-文本"的矩阵,通过对矩阵进行奇异值分解(Singular Value Decomposition, SVD),把文本映射到低维语义空间。矩阵的每一个奇异值及其特征向量代表

* **收稿日期**:2012-12-18
基金项目:国家自然科学基金资助项目(60903225);湖南省自然科学基金项目(11JJ5044);国防科技大学优秀研究生创新基金项目(S100502)
作者简介:谭文堂(1983—),男,贵州平塘人,博士研究生,E-mail:dean.tanw@gmail.com;
肖卫东(通信作者),男,教授,博士,博士生导师, E-mail:wilsonshaw@vip.sina.com

一个潜在话题或者语义维度、概念。LSI 的问题是它无法解释矩阵中负值的物理意义,因此,Hoffman 随后提出概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)[7]。PLSA 把文本与词之间的语义关联表示为概率,PLSA 与 LSI 相比有很大的改进,但它也存在无法处理训练集中未出现的文本和词的问题,而且参数数量随着文本数量呈线性增长。随着概率图模型的发展,Latent Dirichlet Allocation (LDA)为代表的话题模型得到了越来越多的重视[8]。话题模型认为文本是话题的混合,话题是词的混合。它们通过数据的共现关系来揭示数据的隐含语义结构,把文本映射到低维的语义或者话题空间来揭示文本与词的统计特征,减轻了人们分析大规模文本的负担。研究人员在 LDA 之后又提出了 PAM(Pachinko Allocation Model)[9]、CTM (Correlated Topic Model)[10]、RTM(Relational Topic Model)[11]等话题模型,分别考虑了话题的层次结构、关联关系、以及文本之间的关系。话题模型基于概率生成模型,能在没有手工标注的情况下快速分析与理解大规模文本。但是当前大部分话题模型都是面向单一文档集,它们只能分析一个文档集的内在语义结构,不适用于跨文档集的比较性文本挖掘,如:时空演化的文本挖掘[13]、跨文化文本挖掘[3]等。

2.1　LDA 模型

Blei 提出的 LDA 是第一个真正的基于生成模型的话题模型,也是最有影响的话题模型。LDA 模型结构清晰,易于实现高效的参数推理。模型如图 1 所示,模型中话题是一组词的概率分布,而文本则是一组话题的随机混合。LDA 中文本的生成过程如下:

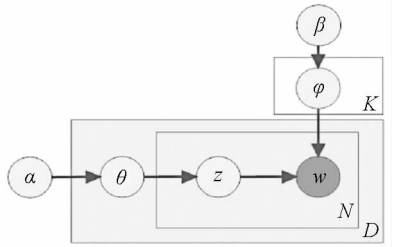

图 1　LDA 模型

Fig. 1　LDA model

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

2)对于每一个文本 $d \in D$,根据 $\theta_d \sim Dir(\alpha)$ 生成其话题的多项式分布参数 θ_d;

3)对于 d 中的第 i 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 $w_{d,i} \sim Mult(\varphi_{z_{d,i}})$ 得到词 $w_{d,i}$。

LDA 之后,David Blei 又提出了动态话题模型(Dynamic Topic Model, DTM)[12]、关联话题模型 CTM。而 Wei Li 等在 2006 年提出了一种层次关联的话题模型 PAM[9],该模型结合了关联话题模型的优点,同时考虑了话题的层次特性。PAM 的优点在于它不仅能对话题的层次关系建模,同时能刻画出话题之间的相关关系。

2.2　跨文本集的话题模型

本节介绍两个主要的跨文本集的比较性话题模型:CCMix 模型和 CCLDA 模型。在此之前先介绍一下本文的相关符号,如表 1 所示。

表 1　符号描述

Tab. 1　Description of symbols

符　号	描　述
α, β, δ	Dirichlet 分布参数
γ_0, γ_1	Beta 分布参数
K	话题数量
z	第 z 个话题
C	文本集数量
d	文本 d
w, t	文本中的一个词 w 或者 t
θ, φ, σ	多项式分布参数
ψ	伯努利分布参数
Dir, Mult, Beta	分别表示狄利克雷分布,多项式分布,贝塔分布

Zhai 等首先考虑了跨文档集的文本挖掘 CTM(Cross-collection Text Mining)[1-2,13-14],他们认为,CTM 主要是发现文档集之间的共同话题以及各文档集对于共同话题的相似与不同之处。CTM 有两个目的,一是发现不同文本集之间的共同讨论的话题,二是分析这些话题在各文档集之间的异同之处。Zhai 等因此提出一种跨文本集的混合模型(Cross-Collection Mixture model, CCMix),该模型实际上是多个 PLSA 的混合:

$$p(w \mid C_i) = (1 - \lambda_B) \times \sum_{j=1}^{k} [\pi_{d,j}(\lambda_C p(w \mid \theta_j) + (1 - \lambda_C) p(w \mid \theta_{j,i}))] + \lambda_B p(w \mid \theta_B) \quad (1)$$

其中 λ_B 代表停用词等噪音的权重,而 λ_C 则体现文本集之间的相似程度。模型采用期望最大法(Expectation-Maximization, EM)求解。该模型基

于 PLSA 模型,虽然简单易于实现,但是其参数数量随着文本增加呈线性增长。

Michael Paul 等则基于 LDA 模型提出了跨文本集的 LDA 模型(Cross-Collection LDA,CCLDA)[3]。如图 2 所示,该模型假设一个话题与两个词的分布关联,其中一个文档集之间共享的,即话题的公共部分,另外一个则是与具体的文本集相关的,即话题在各文本集体现出来的差异。

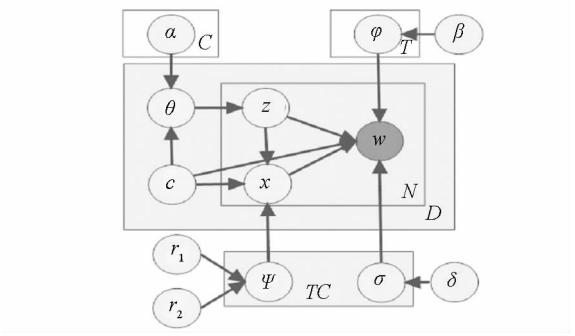

图 2 CCLDA 模型

Fig. 2 CCLDA model

1) 对于每个话题 z 和文本集 c,根据 Beta 分布 Beta(γ^0,γ^1) 获得该话题的一个与文本集有关的伯努利分布向量 ψ_c^z。

2) 对于文本集 c 中的文本 d,根据 Dirichlet 分布 $Dir(\alpha)$ 获得该文本的一个话题的分布向量 θ^d。对于文本 d 中的每一个词 w_i:

a) 根据 θ^d 获取一个话题 z_i

b) 根据 ψ_c^z 抽样得到 x_i

c) 如果 $x_i = 0$,则根据 φ^z 抽样得到一个词作为 w_i;如果 $x_i = 1$,则根据 σ_c^z 抽样得到一个词作为 w_i。

该模型在对词进行抽样时多了一个掷硬币的过程,掷硬币的结果决定该词从话题的哪一个分布生成。这个概率由 γ^0 与 γ^1 决定。该模型的缺点在于会把一些不相干的话题放在一起,特别是当旧的话题消失或者新的话题出现时,话题的可解释性较低。在文献[15]中,Paul 等基于 CCLDA 模型提出一种多方面的话题模型 TAM(Topic-Aspect Model),同时分析多文本集的主题方面和话题。跨文本集的文本挖掘的另一个方面的研究集中在比较性文本摘要[16-18],Kim 等通过一个优化框架对意见相左的多个文本集进行摘要[17],Michael Paul 等则基于 TAM 模型对观点相对的多文本集进行摘要,文献[18]则在 TAM 的基础上提出一种跨文本集的话题 - 方面模型(Cross-collection Topic-Aspect Model, CCTAM) 并用于新闻事件和社会媒体流的摘要。

本文认为同一话题在不同时期,词的分布可能不一样,如“我爸是李刚”,在“李启铭撞人致死”案件中,腐败、特权、富二代等词就常与“李刚”一起出现。而在杭州飙车案中,富二代则与飙车共现。当前主要的话题模型都无法抓住这样的涌现现象,Gabriel Doyle 在 2009 年使用狄利克雷合成多项式分布替代 LDA 模型中的多项式分布得到 DCMLDA(Dirichlet Compound Multinomial LDA) 模型[20],如图 3 所示,但是该模型也是针对单文本集的,而且 ϕ 由于文本的稀疏而难于解释。本文利用 DCM 涌现的特性,把 DCM 模型应用到比较性文本挖掘中。

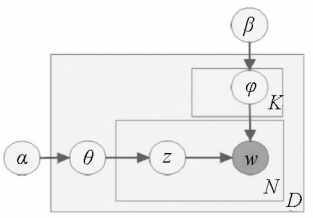

图 3 DCMLDA 模型

Fig. 3 DCMLDA model

3 CDCMLDA 模型

文本模型一般采用多项式分布来对停用词、公用词等进行建模,但是对于在某些文本里涌现的词,多项式分布的效果并不好,因此,Madsen 提出用 DCM 分布来对文本建模。DCM 是一个复合分布,最早由 Minka 在 2003 年提出[21],其中的多项式分布的先验为狄利克雷分布,因此,DCM 也叫玻利亚分布(Polya Distribution),其概率分布为如下所示:

$$p(x \mid \alpha) = \int p(x \mid p)p(p \mid \alpha)\,\mathrm{d}p = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k n_k + \alpha_k)} \prod_k \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \quad (2)$$

对于一个文本而言,向量 α 的和越小,则一个词涌现的概率就越高,当 α 的和为无穷大时,DCM 模型与一个多项式分布等价。假设把文本看作一个词袋,则 DCM 则可以认为是一个装了多个词袋的袋子,DCM 模型的优势在于它能利用狄利克雷分布的聚类特性来对词的涌现进行建模。实验证明,DCM 比传统的多项式要好[5]。

鉴于比较性文本挖掘常常面对地域、文化、时间上涌现的话题,因此,本文使用 DCM 模型来对文本集的词的涌现现象进行建模,假设一个词一旦在一个文档集里出现,则其在该文档集里再次出现的概率更高。同一话题在不同的文档集里,其

词的分布是不一样的,即 φ 在不同的文本集里是有差别的。因此,本文在生成一个文本时,首先生成一个文本集相关的话题与词的分布,再根据该分布生成文本集里的文本。根据该假设,本文提出一种跨文本集的可比话题模型 CDCMLDA,该模型的概率图如图 4 所示,根据上述假设,CDCMLDA 模型的生成过程如下:

1) 对于每个文本集 c 中的每个话题 $k \in \{1, \cdots, K\}$,根据 $\phi_{kc} \sim Dir(\beta_k)$ 抽样得到话题与词的分布;

2) 对于文本集 c 中的每个文本 $d \in \{1, \cdots, D\}$,根据 $\theta_{cd} \sim Dir(\alpha)$ 得到文本中话题的分布;

3) 对于文本 d 中的每个词 $n \in \{1, \cdots, N_{cd}\}$:

a) 根据 $z_{cdn} \sim \theta_{cd}$ 获得一个话题 z_{cdn};

b) 根据 $w_{cdn} \sim \phi_{z_{cdn}d}$ 获得一个词当作 n。

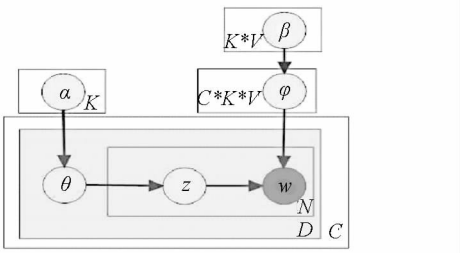

图 4　CDCMLDA 模型

Fig. 4　CDCMLDA model

由上述生成过程,设 C 为所有文本集的集合,第 c 个文本集的文本数量为 M_c,CDCMLDA 模型的联合分布函数为:

$$p(C) = \prod_{c=1}^{C}\prod_{m=1}^{M_c} p(\boldsymbol{w}_{cm}, \boldsymbol{z}_{cm}, \boldsymbol{\theta}_{cm}, \boldsymbol{\Phi} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{c=1}^{C}\prod_{m=1}^{M_c}\prod_{n=1}^{N_m} p(w_{c,m,n} \mid \boldsymbol{\varphi}_{z_{c,m,n}}) \times p(z_{c,m,n} \mid \boldsymbol{\theta}_{cm}) \times p(\boldsymbol{\theta}_{cm} \mid \boldsymbol{\alpha}) \times p(\boldsymbol{\Phi}_c \mid \boldsymbol{\beta}) \tag{3}$$

CDCMLDA 与 DCMLDA 模型都是基于 DCM 来对文本中词的涌现进行建模,但是本文模型具有以下不同:首先,文本模型针对多个可比的文本集,而 DCMLDA 模型面向单文本集;其次本文模型的局部话题是文本集内的词的分布,φ 是可解释的实际话题,而在 DCMLDA 模型中,φ 体现的是一个文本的话题,由于话题是词在多个文本中共现的模式,因此并不具有实际意义,β 参数才是真正的话题分布,因此,CDCMLDA 与 DCMLDA 模型是两个不同的模型,虽然他们都以 DCM 为基础。相对于 DCMLDA,CDCMLDA 从多个文本集的角度出发,侧重文本集之间的可比性。相对于 CCMix 和 CCLDA 模型,CDCMLDA 更加符合文本集的词的涌现规律。在 CCLDA 模型中,作者假设词来自两种话题,一种是在所有文本集中都存在的背景话题。另一种是文本集相关的话题,这两种话题同时存在,通过一个伯努利分布来选择,但在 CCLDA 模型中,该分布在每个文本上都是一样的,即对于一个词 w,$x_i = 0$ 的概率在每个文本中都是一样的,这显然是不对的。另外,不是所有话题都存在背景模型,根据词的涌现现象,话题在一个文本集里出现,在下一个时间段可能消失,代之而起的是另外的话题,因此,并不是所有话题都是可比的。因此,CCLDA 是更倾向于平行比较的话题模型,而 CDCMLDA 则两者兼有之,既有平行的比较,也有纵向的演化。CDCMLDA 的一个更大的不同在于,ϕ 是一个三维变量($C \times V \times K$),CDCMLDA 比 LDA 多了一个 K 维的自由度来描述词在文本集上的涌现现象。

对于复杂的概率图模型,精确的参数推导是不可行的,一般都采用近似推导方法,Blei 在 LDA 模型中采用变分法(Variational EM Algorithm)[8],Griffths 和 Steyvers 则采用 Gibbs 抽样来推导 LDA 模型的参数[22],Gibbs 抽样是一种马尔可夫链蒙特卡洛方法(Markov Chain Monte Carlo Algorithm, MCMC),该方法易于实现,运行效率高,不像变分 EM 方法那样容易陷入局部最优,本文采用 Gibbs 抽样来推导模型参数,因此本文在文本集内部使用 Gibbs 抽样,在文本集层次上,使用EM算法(Expectation-Maximization,EM)来估计超参数α与β,类似于DCMLDA模型的蒙特卡罗期望最大法(Monte Carlo Expectation Maximization),根据话 题与词的关系可得:

$$p(\boldsymbol{w}, \boldsymbol{z} \mid \alpha, \beta) = p(\boldsymbol{w} \mid \boldsymbol{z}, \beta) p(\boldsymbol{z} \mid \alpha) \tag{4}$$

其中,$p(\boldsymbol{w} \mid \boldsymbol{z}, \beta)$ 可表示为式(9),n_{ctz} 表示在文本集 c 中词 t 分配给话题 z 得次数。

$$p(\boldsymbol{w} \mid \boldsymbol{z}, \beta) = \int_{\phi} p(Z \mid \varphi) p(\varphi \mid \beta) \mathrm{d}\varphi = \int_{\phi} p(\varphi \mid \beta) \prod_{c}\prod_{d}\prod_{n=1}^{N_d} \varphi_{c, z_{d,n}} \mathrm{d}\varphi = \int_{\phi} \Big[\prod_{c,z} \frac{1}{B(\beta_{\cdot z})} \prod_{t} (\varphi_{c,t,z})^{\beta_{tz}-1}\Big] \prod_{c,z} (\varphi_{c,t,z})^{n_{ctz}} \mathrm{d}\varphi \tag{5}$$

根据式(5),对于文本集 c 中的文本,$p_c(w \mid z, \beta)$ 可表示为:

$$p_c(\boldsymbol{w} \mid \boldsymbol{z}, \beta) = \prod_{z=1}^{K} \frac{B(n_{cz} + \boldsymbol{\beta}_{\cdot z})}{B(\boldsymbol{\beta}_{\cdot z})} \tag{6}$$

其中 $B(\boldsymbol{n}) = \dfrac{\prod_i \Gamma(n_i)}{\Gamma(\sum_i n_i)}$,$\boldsymbol{\beta}_{\cdot z}$ 表示一个向量。$p_c(z \mid$

$\boldsymbol{\alpha}$) 可表示为:

$$p_c(\boldsymbol{z} \mid \boldsymbol{\alpha}) = \int p(\boldsymbol{z} \mid \theta)\, p(\theta \mid \alpha_c) d\theta = \prod_{m=1}^{M_c} \frac{B(\boldsymbol{n}_m + \boldsymbol{\alpha}_c)}{B(\boldsymbol{\alpha}_c)} \tag{7}$$

根据式(6) 和(7),可得

$$p_c(\boldsymbol{z},\boldsymbol{w} \mid \boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{z=1}^{K} \frac{B(n_{cz} + \boldsymbol{\beta}_z)}{B(\boldsymbol{\beta}_z)} \cdot \prod_{m=1}^{M_c} \frac{B(\boldsymbol{n}_m + \boldsymbol{\alpha}_c)}{B(\boldsymbol{\alpha}_c)} \tag{8}$$

根据链式规则,设 $\boldsymbol{w} = \{w_i, \boldsymbol{w}_{\neg i}\}$, $\boldsymbol{z} = \{z_i = k, \boldsymbol{z}_{\neg i}\}$, Gibbs 抽样的迭代式为:

$$\begin{aligned} p_c(z_i = k \mid \boldsymbol{z}_{\neg i}, \boldsymbol{w}) &= \frac{p(\boldsymbol{w},\boldsymbol{z})}{p(\boldsymbol{w},\boldsymbol{z}_{\neg i})} \\ &= \frac{p(\boldsymbol{w} \mid \boldsymbol{z})}{p(\boldsymbol{w}_{\neg i} \mid \boldsymbol{z}_{\neg i})} \cdot \frac{p(\boldsymbol{z})}{p(\boldsymbol{z}_{\neg i})} \\ &\propto \frac{B(n_{cz} + \boldsymbol{\beta}_z)}{B(n_{cz,\neg i} + \boldsymbol{\beta}_z)} \cdot \frac{B(\boldsymbol{n}_m + \boldsymbol{\alpha}_c)}{B(\boldsymbol{n}_{m,\neg i} + \boldsymbol{\alpha}_c)} \\ &= \frac{n_{ctz} + \beta_{tz}}{\sum_t n_{ctz} + \beta_{tz}} \cdot \frac{n_{cmz} + \alpha_{cz}}{\sum_k n_{cmz} + \alpha_{cz} - 1} \end{aligned} \tag{9}$$

根据 Dirichlet 分布的期望,隐含变量 φ 与 θ 计算公式为:

$$\begin{aligned} \varphi_{czt} &= \frac{n_{czt} + \beta_{zt}}{\sum_t n_{czt} + \beta_{zt}} \\ \theta_{cmz} &= \frac{n_{czm} + \alpha_{cz}}{\sum_z n_{czm} + \alpha_{cz}} \end{aligned} \tag{10}$$

在很多 LDA 的应用中,α 和 β 都是人为设定,根据 Grffiths 和 Steyvers 的建议[22], 一般 $\alpha = 50/K$, $\beta = 0.1$,LDA 模型即可得到较好的效果,但是本文中的 α 和 β 参数具有不同的意义,因此本文采用与 Gabriel Doyle 类似的方法,使用 Single-Sample Monte Carlo EM算法来估计模型的参数,该方法通过最大化 $p(\boldsymbol{z},\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ 来求解 α 和 β,但是在期望最大化的每一步采用 Gibbs 抽样来推导其他参数。把 $p(\boldsymbol{z},\boldsymbol{w} \mid \boldsymbol{\alpha},\boldsymbol{\beta})$ 中的 Beta 函数展开可得:

$$\begin{aligned} p(\boldsymbol{z},\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_c^C \Big[\prod_{z=1}^{K} \frac{B(\boldsymbol{n}_{cz} + \boldsymbol{\beta}_z)}{B(\boldsymbol{\beta}_z)} \cdot \prod_{m=1}^{M_c} \frac{B(\boldsymbol{n}_m + \boldsymbol{\alpha}_c)}{B(\boldsymbol{\alpha}_c)} \Big] \\ &= \prod_c^C \prod_m^{M_c} \frac{(\prod_z \Gamma(n_{czm} + \alpha_{cz}))\Gamma(\sum_z \alpha_{cz})}{(\prod_z \Gamma(\boldsymbol{\alpha}_{cz}))\Gamma(\sum_z \boldsymbol{n}_{czm} + \boldsymbol{\alpha}_{cz})} \\ &\times \prod_c^C \prod_z^K \frac{(\prod_t \Gamma(n_{czt} + \beta_{t,z}))\Gamma(\sum_t \beta_{t,z})}{(\prod_t \Gamma(\beta_{t,z}))\Gamma(\sum_t n_{czt} + \beta_{t,z})} \end{aligned} \tag{11}$$

通过对对数似然函数进行最大化,将 $p(\boldsymbol{z},\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ 取对数:

$$\begin{aligned} L(\alpha, \beta; w, z) &= \sum_{c,m,z} [\ln\Gamma(n_{czm} + \alpha_{cz}) - \ln\Gamma(\alpha_{cz})] \\ &+ \sum_{c,z} [\ln\Gamma(\sum_z \alpha_{cz}) - \ln\Gamma(\sum_z \boldsymbol{n}_{czm} + \boldsymbol{\alpha}_{cz})] \\ &+ \sum_{c,z,t} [\ln\Gamma(n_{czt} + \beta_{t,z}) - \ln\Gamma(\beta_{t,z})] \\ &+ \sum_{c,z} [\ln\Gamma(\sum_t \beta_{t,z}) - \ln\Gamma(\sum_t n_{czt} + \beta_{t,z})] \end{aligned} \tag{12}$$

(12) 式中的前两项只与 α 相关,后两项只与 β 相关,因此可以对两者分别最大化:

$$\begin{aligned} \boldsymbol{\alpha} = \operatorname{argmax} &\sum_{c,m,z} [\ln\Gamma(n_{czm} + \alpha_{cz}) - \ln\Gamma(\alpha_{cz})] \\ &+ \sum_{c,z} [\ln\Gamma(\sum_z \alpha_{cz}) - \ln\Gamma(\sum_z \boldsymbol{n}_{czm} + \boldsymbol{\alpha}_{cz})] \end{aligned} \tag{13}$$

$$\begin{aligned} \boldsymbol{\beta}_z = \operatorname{argmax} &\sum_{c,z,t} [\ln\Gamma(n_{czt} + \beta_{t,z}) - \ln\Gamma(\beta_{t,z})] \\ &+ \sum_{c,z} [\ln\Gamma(\sum_t \beta_{t,z}) - \ln\Gamma(\sum_t n_{czt} + \beta_{t,z})] \end{aligned} \tag{14}$$

通过上述 $K+1$ 个约束优化问题,可以得到 α 和 β 的值,本文采用有限记忆的 BFGS 方法来求解上述最大化问题[23]。

关于 β 的取值,本文还采取了一种基于 LDA 模型的方法,由于 β 实际上就是在所有文本上话题对词的分布,因此,本文首先使用 LDA 模型在所有文本上进行训练得到 φ 之后,把 φ 作为 CDCMLDA 模型的 β 来进行参数推导,这样可以省掉大量的计算。

4 实验和分析

4.1 训练数据

本文实验的目的在于验证 CDCMLDA 模型的性能、效率,我们选用以下数据集,分别是 NIPS 会议数据集、Michael Paul 的 tourists 数据集[3] 和 Qiaozhu Mei 整理的有关印度尼西亚海啸 Tsunami 数据集[14]。

表 2 数据集描述

Tab. 2 Description of experimental data

数据集名称	文本数	时间、地域跨度
Tsunami	7468	12/19/04 – 02/08/05
NIPS	238	2001 – 2005
Tourists	2461	英国、印度、新加坡

三个数据集的属性信息如表2 所示。Tsunami 是 Qiaozhu Mei 整理的有关印度洋海啸的新闻报

道,分别来自 BBC,CNN,路透社、新华社、纽约时报、印度时报等多家著名媒体的报道。NIPS 会议数据集包含了2001 年 ~ 2005 年的所有 NIPS会议的论文全文。Tourists 数据集则是 Michael Paul 作跨文化分析时搜集的印度、新加坡、英国三个国家有关旅游的博客数据。试验数据将分别从三个数据集中选取。实验分为四个部分:面向新闻事件演化的话题分析,主要针对 Asia Tsunami 数据集;学术研究的演化分析,主要针对 NIPS 数据集;地域比较性分析,主要针对 tourists 数据集;模型困度(perplexity) 指标分析[8]。

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{m=1}^{M}\log p(w_m)}{\sum_{m=1}^{M}N_m}\right\} \quad (15)$$

其中 $p(w_m)$ 为:

$$\begin{aligned} p(w_m \mid M) &= \prod_{n=1}^{N_m}\sum_{k=1}^{K}p(w_n = t \mid z_n = k) \\ &\quad \cdot p(z_n = k \mid d = m) \\ &= \prod_{t=1}^{V}\{\sum_{k=1}^{K}\varphi_{k,t}\cdot\theta_{m,k}\}^{n_m^t} \end{aligned} \quad (16)$$

困惑度是当前最常用的度量文本模型预测能力的指标,检验模型基于当前训练结果预测未来文本的能力,该指标可解释为模型生成测试数据所需要的均匀分布的词的个数的期望,意味着模型体现隐含语义结构的能力,困惑度越低,说明模型建模效果越好[6,17]。

4.2　实验结果

4.2.1　面向新闻事件演化的话题分析

本文从 Asia Tsunami 数据集中选择路透社在 2004 年 ~2005 年的报道,利用 CDCMLDA 模型分析路透社的报道在两者之间的差异。数据包含 1565 篇报道,本文把数据划分为 2004 年和 2005 年两部分。实验结果如表 3 所示,表中的数值表示词在该话题中的权重,该数值限于篇幅做了省略。从表 3 中我们发现路透社在 2004 年的有关报道是在事件发生不久,根据当时的现状做出的报道,而 2005 年则回顾性地结合实时新闻进行报道。在话题 1 中,2004 年的报道偏向英国对于海啸的反映以及做出的援助和拨款,2005 年则偏向海啸对周边国家的影响、规模和国际社会的援助。话题 2 主要是讨论人员的伤亡和各国政府的反映,从 2004 年可以看出,当时媒体的报道重点在于伤亡人员的搜救工作和各国政府对海啸采取的行动,而 2005 年则更偏向总结地震本身带来的人员伤亡和纪念活动。话题 3 则讨论当地的生活状况,2004 年的报道侧重于灾难发生后的食物与水的供给、伤病、交通等,2005 年则侧重于重建的情况和人们的生活。

表 3　路透社在 Asia Tsunami 事件上的话题变迁

Tab. 3　Topic evolution on Asia Tsunami reports of Reuters

话题 1		话题 2		话题 3	
nations countries debtbillion		British Foreign missing government		water city Indonesia aid food quake	
Annan aid tsunami relief world		confirmed dead told disaster killed		supplies northern streets airport	
2004	2005	2004	2005	2004	2005
Nations 0.0463	tsunami 0.06		disaster 0.0298	coast 0.027	water 0.0415
debt 0.03173	countries 0.0342	global 0.04	victims 0.0272	survivors 0.0273	coast 0.027556
Brown 0.0194	Indonesia 0.031	warning 0.033	global 0.0239	water 0.0262	tsunami 0.0221
countries 0.0191	earthquake 0.02	effort 0.0302	million 0.023	food 0.023	people 0.0187
meeting 0.0156	Reuters 0.02455	leaders0.026	world 0.0217	said 0.021	ground 0.0171
pounds 0.01548	hit 0.02363	pledged 0.025	celebrations 0.02	city 0.021	area 0.0165
Bank 0.013692	continued 0.02	continued 0.0231	urged 0.0138	capital 0.017	wave 0.0158
reconstruction	waves 0.01964	help 0.0210	Britain 0.01316	Indonesian	quake 0.015
0.01369	Indian 0.016983	humanitarian 0.02	give 0.01263	0.017	relatives 0.013
Germany 0.013	Ocean0.0141	Aid 0.019	called 0.01224	Sumatra 0.015	fishermen 0.012
Blair 0.012	cost 0.0126	efforts 0.017	silence 0.01224	airport 0.015	miles 0.0119
Billion 0.012	triggered 0.011	But 0.015	millions 0.0122	lost 0.0149	town 0.0108
Britain 0.012	catastrophe 0.01	Indian 0.014	asked 0.0117	quake 0.0134	swept 0.0103
moratorium 0.01	system 0.01	millions 0.0141	site 0.0111	area 0.0133	debris 0.0102
aid 0.01113	magnitude 0.01	including 0.012	Minister 0.0111	injured 0.0123	Madras 0.0091
world 0.011	powerful 0.0104	United 0.0127	cancelled 0.0111	trying 0.0108	sea 0.0089
Minister 0.010	nations 0.009	conference 0.01	city 0.010529	tour 0.0105	inland 0.0088
plan 0.00985	largest 0.0094	lives 0.0117	Asia 0.0103	destroyed 0.010	night 0.008
Prime 0.009725	economic 0.007		children 0.0101	supplies 0.009	flooded 0.008

表4 NIPS 2003 – 2004 话题变迁

Tab. 4 Topic evolution on proceeding of NIPS between 2003 and 2004

话题1		话题2		话题3	
Learning state predictions Action prediction option time observation agent history probability sequence environment tests algorithm current		points data distance manifold local space dimensionality projection embedding nearest neighbor learning reduction Euclidean linear algorithms		Model image Figure human motion Representation object pose Vision Detection surface video similar texture tracking visual segmentation	
2003	2004	2003	2004	2003	2004
training 0.081	feature 0.0709	distance 0.075	dimensional0.032	image 0.031	model 0.0604
test 0.0391	regression 0.067	dimension 0.044	components0.021	object 0.013	image 0.0232
prediction0.037	space 0.0642	method 0.036	space 0.021	model 0.01	human 0.0126
examples 0.032	matrices 0.04	manifold 0.035	local 0.019	face 0.0094	This 0.0074
instance 0.0315	data 0.0313	reduction 0.027	point 0.019	detection 0.0074	latent 0.0074
Learning 0.017	factors 0.02	data 0.0236	vectors 0.018	appearance 0.0069	motion 0.0068
SVMs 0.0158	dimensional0.019	methods 0.0226	analysis 0.0156	features 0.0069	Figure 0.0058
hand 0.01516	weighted 0.018	nearest 0.02201	distance 0.0153	pixels 0.0065	parameters0.0057
threshold 0.014	infinite 0.0174	database0.01834	index 0.0147	pose 0.0065	data 0.0057
accuracy 0.014	variable 0.012	information0.017	projection 0.0138	recognition 0.0061	inference 0.0056
machine0.0136	observation 0.012	dataset 0.016	manifold 0.0135	local 0.006	object 0.0053
context 0.0124	vector 0.0115	mass 0.0155	neighbors 0.012	segmentation0.005	generative0.0053
synthetic 0.011	mapping 0.0107	estimate 0.014	reduction 0.0119	tracking 0.0057	range 0.0052
extended 0.010	synthetic 0.0105	interactions0.012	hand 0.0112	patches 0.0054	learned 0.005
original 0.0105	scaling 0.0097	types 0.012	vertices 0.01013	Vision 0.0048	structure 0.0046
achieve 0.0103	Lasso 0.009	motif 0.011	greedy 0.0097	single 0.0048	mixture 0.0044
noisy 0.0101	coordinates0.008	amino 0.01	algorithms 0.009	surface 0.0046	level 0.0042
boundary0.008	predictions0.0087	plane 0.01	principal 0.0092	background0.003	hierarchical0.003

4.2.2 学术研究的演化分析

本文利用 CDCMLDA 模型来分析 NIPS 著名机器学习方面的会议,使用 NIPS 会议 2003 年 ~ 2004 年的论文集。实验结果如表 4 所示,限于篇幅,本文只选取了其中三个话题,话题 1 与监督学习研究相关,通过表 4 可以看出,在 2003 年,NIPS 的文章在监督学习方面侧重于 SVM 等方法的有关理论,而在 NIPS2004 中,更偏向特征选择、面向特征选择的降维方法等。话题 2 则是有关流形学习、数据降维方面的研究,NIPS2003 和 2004 都关注了这个方面的研究,二者之间区别不大,但 2003 年倾向于流形学习在高维数据中的应用,而 2004 年则更关注流形学习的算法本身。话题 3 是有关图像、视频方面的研究。NIPS2003 关注人脸、动作、表情的识别,背景模型的抽取等;NIPS2004 则更关注生成模型、层次贝叶斯理论在图像、视频中的应用。

4.2.3 地域比较性分析

Tourists 数据集是 Michael Paul 等从 lonelyplanet.com 搜集的有关旅游的博客,本文从中抽取部分作为实验数据,包括已经去过或者准备去英国和新加坡两地旅游的用户所写的博客。博客包含了用户在该地区的旅游见闻,以及准备去旅游的用户关心的一些问题。实验结果如表 5 所示。本文选择其中三个话题作为代表,话题 1 主要讨论了旅游的交通工具,两个文本集分别从伦敦和新加坡的角度介绍了两地出行的交通问题。话题 2 则讨论了两地旅游的多个方面,著名景点、音乐等艺术、街道、食物等。从话题 2 可看出,英国游客游览的更多是名胜古迹,如博物馆等,还关注了英国的音乐表演。而到新加坡的游客关注一些花园、游乐场、酒吧等娱乐场所。在话题 3 中两地游客都讨论了在两地旅游的支付问题,包括信用卡、小费等方面。

表 5　有关伦敦与新加坡的博客话题比较

Tab. 5　The comparison between the Blogs about London and Singapore

话题 1		话题 2		话题 3	
Airport train bus get take hours flight check station luggage taxi Terminal ticket leave		Good place great street area nice road Restaurants food walk Pubs park music old shopping house bars		Card know get bank pay money credit phone using cash account fee charge free service check	
London	Singapore	London	Singapore	London	Singapore
heathrow0.015	transit 0.013	city 0.0145	place 0.021	card 0.020	card 0.031
London0.0149	city 0.0121	great 0.0144	great 0.013	use 0.0192	use 0.024
travel 0.01	free 0.0093	pubs 0.01254	area 0.013	bank 0.018	know 0.019
station 0.0099	immigration0.009	nice 0.0124	orchard 0.01	know 0.016	money 0.017
hours 0.0092	singapore 0.0084	place 0.0116	street 0.01	get 0.015	get 0.0144
tickets 0.009	hour 0.0076	music 0.010	top 0.01	pay 0.013	pay 0.0135
day 0.0082	leave 0.0069	area 0.0093	food 0.009	money 0.013	bank 0.012
flybudget0.007	morning 0.0068	food 0.0088	shopping0.009	credit 0.0127	credit 0.012
book 0.007	customs 0.0059	old 0.0086	nice 0.007	phone 0.012	cash 0.0112
guide 0.007	johore 0.0057	centre 0.0084	style 0.007	number 0.012	rate 0.009
frills 0.007	catch 0.0053	museum0.0068	quay 0.006	account 0.011	service 0.009
global 0.0067	tour 0.0048	town 0.0064	centre 0.006	charge 0.0099	buy 0.008
luggage 0.006	arrive 0.0047	south 0.0059	building 0.006	fee 0.0090	atm 0.008
journey 0.006	clear 0.0046	tourists 0.0057	plenty 0.006	cash 0.0089	charge 0.008
cost 0.006	going 0.0045	gardens 0.0056	bars 0.006	free 0.0088	account 0.007
advance 0.006	town 0.0044	find 0.00539	park 0.0058	give 0.0081	safe 0.007
cheaper 0.005	cab 0.0044	hill 0.0053	house 0.0056	using 0.0079	local 0.006
rail 0.005	shuttle 0.0044	top 0.0053	beer 0.0053	mobile 0.0073	fee 0.006
express 0.005	baggage 0.0043	bridge 0.0052	corner 0.0053	find 0.0069	exchange 0.0067
service 0.004	fare 0.0042	river 0.0051	garden 0.0049	check 0.0069	countries 0.006

为了对比 CDCMLDA 模型和 CCLDA 模型,本文用 CCLDA 模型对 Tourists 数据集进行分析。表 6 列出了与表 5 相近的三个话题。从表 5 与表 6 中可看出,在前两个话题上,两个模型的差别不大,但在话题 3 上,CDCMLDA 模型的结果显然更具可解释性。

4.2.4　模型困惑度评估

本节对 CDCMLDA 模型进行定量的评估,通过比较 CCLDA、LDA 和 CDCMLDA 模型的模型困惑度来评估模型的预测能力。试验在 Tourists 数据集和 NIPS 数据集上进行。实验采用数据集的 80% 作为训练数据,20% 作为测试数据,计算模型的模型困惑度(Perplexity)指标。试验中,三个模型都对训练数据进行 2000 次采样,实验结果如图 5 和图 6 所示。CDCMLDA 模型在两个数据集上的模型困惑度相比 CCLDA 和 LDA 模型小很多,这表明 CDCMLDA 模型赋予最可能在文本出现的词以较高的概率,模型具有较好的预测能力。

5　结论与下一步工作

本文针对动态、可比的文本集提出一种面向比较性文本挖掘的话题模型 CDCMLDA,该模型首先假设同一个话题在不同时期或者不同文本集之间具有差异,采用 DCM 模型来对不同文本集之间词的涌现现象建模,基于 DCM 建立一种跨文本集的话题模型 CDCMLDA;其次该模型采用 Monte Carlo EM 算法来推导模型参数,在文本集内部使用 Gibbs 抽样,在文本集之间用 EM 算法实现超参数的估计。实验证明,在多个数据集上 CDCMLDA 模型能够发现同一话题之间的差异,对词在不同文本集之间的涌现进行建模。实验结果显示,模型所得话题具有较好的可解释性。在

表6 有关伦敦与新加坡博客话题 CCLDA 模型分析结果

Tab. 6 The comparison between the Blogs about London and Singapore by CCLDA

话题1		话题2		话题3	
airport check time know airlines com hotel take flights don fly station cheap cost book flying hour journey		good food want where know think area time park nice free bit transport stuff places street old night		www stay cheap good night try budget hostel look london cheaper card etc website live search price	
London	Singapore	London	Singapore	London	Singapore
london 0.081	airport 0.060	london 0.045	food 0.032	com 0.045	good 0.057
heathrow 0.044	changi 0.050	saturday 0.024	mrt 0.032	london 0.031	place 0.036
www 0.038	flight 0.036	cold 0.020	india 0.029	hotel 0.019	find 0.033
train 0.037	taxi 0.034	talk 0.020	night 0.029	know 0.017	want 0.032
global 0.030	terminal 0.026	sitting 0.020	zoo 0.024	looking 0.017	where 0.024
flybudget 0.029	budget 0.025	baby 0.019	road 0.020	get 0.016	time 0.019
frills 0.029	city 0.0256	grey 0.019	sentosa 0.019	bus 0.013	com 0.019
guide 0.027	transit 0.025	dream 0.019	shopping 0.019	find 0.013	bit 0.018
manchester	mrt 0.021	lift 0.018	orchard 0.018	free 0.011	great 0.01
0.016	air 0.020	comin 0.018	chinatown 0.017	great 0.0105	lot 0.015
airlines 0.015	hours 0.019	skirt 0.018	centre 0.015	worth 0.010	money 0.015
cambridge 0.012	time 0.019	heels 0.018	bugis 0.013	birmingham	going 0.014
faq 0.0128	tiger 0.018	drinkers 0.018	quay 0.013	0.010	years 0.013
lot 0.012	free 0.018	deals 0.017	eat 0.013	pay 0.0097	look 0.013
stansted 0.011	immigration	acid 0.017	great 0.013	accommodation	expensive 0.012
bristol 0.011	0.0159	screaming 0.017	try 0.012	0.009	things 0.011
ryanair 0.010	asia 0.015	girl 0.017	cheap 0.012	expensive 0.009	couple 0.011
gatwick 0.010	airline 0.013	station 0.017	city 0.012	central 0.008	maybe 0.011
pounds 0.010	luggage 0.013	door 0.017	safari 0.011	house 0.007	pretty 0.010
darwin 0.010	leave 0.0119	tube 0.017	walk 0.011	room 0.007	available 0.009
snow 0.0100	singapore 0.011	pubs 0.013	good 0.010		cost 0.008

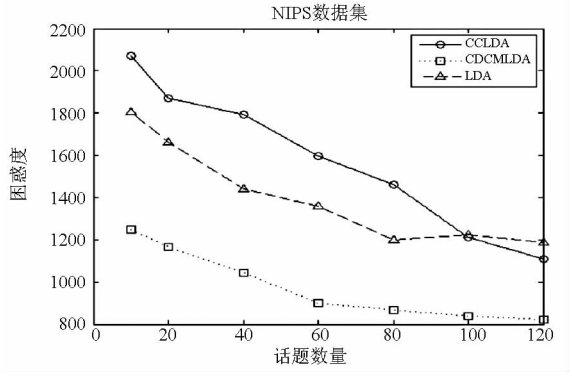

图5 NIPS 数据集上三个模型困惑度比较

Fig. 5 The comparison of perplexity of three models on dataset NIPS

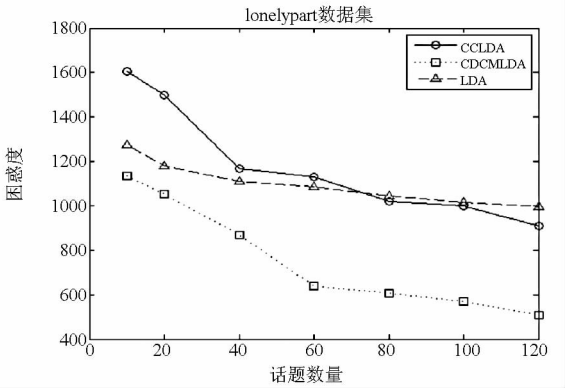

图6 Tourists 数据集上三个模型困惑度比较

Fig. 6 The comparison of perplexity of three models on dataset Tourists

模型困惑度方面,CDCMLDA 模型相对于 LDA、CCLDA 模型具有明显的优势。

下一步的工作在于优化模型参数估计的算法,使用变分法来实现参数的快速推导,在此基础上建立非参数化的 CDCMLDA 模型,并考虑把 CDCMLDA 模型应用到跨文本集的自动摘要方面。

致谢 感谢斯坦福大学的 Gabriel Doyle 博士关于实现的建议和帮助。

参考文献(References)

[1] Zhai C, Atulya V, Bei Y. A cross-collection mixture mode for comparative text mining[C]//Proceedings of The International Conference on Knowledge Discovery and Data Mining. Seattle, Washington,USA: ACM, 2004: 743 - 748.

[2] Yin Z, Cao L,Jiawei Han, et al. Geographical topic discovery and comparison [C]//Proceedings of The International Conference on World Wide Web. Hyderabad, India, 2011: 247 - 256.

[3] Paul M, Girju R. Cross-cultural analysis of blogs and forums with mixed-collection topic models [C]//Proceedings of The 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 2009:1408 - 1417.

[4] Paul M,Girju R. Comparative scientific research analysis with a language-independent cross-collection model[C]//Proceedings of SEPLN , Valencia, Spain,2010:153 - 160.

[5] Madsen R E, Kauchak D, Elkan C. Modeling word burstiness using the dirichlet distribution[C]//Proceedings of the International Conference on Machine Learning, New York: ACM, 2005: 545 - 552.

[6] Deerwester S, Dumais S, Furnas G, et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science, 1990:41:17.

[7] Hofmann T. Probabilistic latent semantic indexing [C]// Proceedings of SIGIR, 1999:50 - 57.

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

[9] Li W, McCallum A. Pachinko allocation: DAG-Structured mixture models of topic correlations [C]//Proceedings of the International Conference on Machine Learning, Pittsburgh, PA,2006: 577 - 584.

[10] Blei D M, Lafferty J D. Correlated topic models [C]// Proceedings of the Advances in Neural Information Processing Systems, 2006.

[11] Chang J, Blei D M. Relational topic models for document networks[C]//Proceedings of the international conference on artificial Intelligence and Statistics, Clearwater Beach, Florida, USA; 2009:81 - 90.

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

[13] Mei Q, Liu Cao, Su Hang, et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs [C]// Proceedings of the International Conference on World Wide Web, Edinburgh, Scotland, 2006: 533 - 542.

[14] Mei Q, Zhai C. Discovering evolutionary theme patterns from text-an exploration of temporal text mining[C]//Proceedings of the International Conference on Knowledge Discovery and Data Mining, Chicago,Illinois,USA, 2005:198 - 207.

[15] Paul M, Girju R. A two-dimensional topic-aspect model for discovering multi-faceted topics[C]//Proceedings of the 24th Conference on Artificial Intelligence, 2010:545 - 550.

[16] Paul M, Zhai C, Girju R. Summarizing contrastive viewpoints in opinionated text[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010: 66 - 76.

[17] Kim H, Zhai C. Generating comparative summaries of contradictory opinions in text [C]//Proceeding of the 18th ACM Conference on Information and Knowledge Management, 2009:385 - 394.

[18] Wang D D, Ogihara M, Li T. Summarizing the differences from Microblogs [C]//Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval, 2012:1147 - 1148.

[19] Gao W, Li P, Darwish K. Joint topic modeling for event summarization across news and social media streams [C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012: 1173 - 1182.

[20] Doyle G, Elkan C. Accounting for burstiness in topic models [C]//Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada, 2009:545 - 552.

[21] Minka T. Estimating a dirichlet distribution [R]. MIT, 2000.

[22] Griffths T L, Steyvers M. Finding scientific topics[J]. PNAS Early Edition, 2004:1 - 8.

[23] Zhu C, Byrd R H, Lu P, et al. Algorithm 778: L-BFGS-B: Fortran routines for large scale bound constrained optimization [J]. ACM Transactions on Mathematical Software, 1997,23: 550 - 560.

[24] Heinrich G. Parameter estimation for text analysis [R]. Fraunhofer Institute for Computer Graphics, 2009.