•可信计算与信息安全•

DOI:10.15961/j.jsuese.201700626

基于改进的ccLDA多数据源热点话题检测模型

陈兴蜀[1,2], 马晨曦[2], 王文贤[3], 高　悦[2], 王海舟[1*]

(1.四川大学 网络空间安全学院, 四川 成都 610065; 2.四川大学 计算机学院, 四川 成都 610065;
3.四川大学 网络空间安全研究院, 四川 成都 610065)

摘　要:目前, 跨文本集的话题发现模型(cross-collection LDA, ccLDA)只适用于各个数据源话题相似度很高的场景, 而且其全局话题和每个数据源的局部话题会强制对齐, 存在词语稀疏的问题。针对ccLDA模型中的不足, 提出了改进的跨文本集话题发现模型(improved ccLDA, IccLDA)。该模型在采样时先判断词语属于全局话题还是局部话题, 再分别进行采样, 避免了ccLDA模型中全局话题和局部话题必须对齐的缺点, 进而降低了词语在全局话题和局部话题的分散程度, 使该模型可以适用于多数据源的场景。在公开数据集上进行了多数据源文本集的话题发现实验, 并进行了话题比较性分析。实验结果表明, 在设置不同的话题数时, IccLDA模型的困惑度值均低于LDA模型和ccLDA模型, 表明IccLDA模型具有更优的建模能力。最后, 在真实数据集上开展了进一步实验验证, 证明了本文提出的改进模型不仅建模能力优于原始模型, 还可以有效地发现各个数据源讨论的公共话题和每个数据源讨论的局部话题, 更适用于多数据源场景的文本话题发现。

关键词:话题检测; 话题模型; LDA; 多数据源; IccLDA

中图分类号:TP391.1　　**文献标志码**:A　　**文章编号**:2096-3246(2018)02-0141-07

Multi-source Topic Detection Analysis Based on Improved ccLDA Model

CHEN Xingshu[1,2], *MA Chenxi*[2], *WANG Wenxian*[3], *GAO Yue*[2], *WANG Haizhou*[1*]

(1.College of Cybersecurity,Sichuan Univ.,Chengdu 610065,China; 2.College of Computer Sci.,Sichuan Univ.,Sichuan,Chengdu 610065,China;
3.Cybersecurity Research Inst.,Sichuan Univ.,Chengdu 610065,China)

Abstract: At present,ccLDA (cross collection LDA) model has been found only applicable to data sources that topic similarity is very high,and its global topics and local topics of each data source will be forced alignment,hence causing words sparse.In order to solve the problem of ccLDA model,an improved ccLDA topic model (IccLDA) was proposed.When sampling,this model firstly decides whether words are global topics or local topics,and then takes samples respectively.In this way,it can avoid the problem that the global topics and local topics in ccLDA model must be aligned,and also can reduce the dispersion degree of the words in the global topics and local topics,making the model suitable for multiple data source scenarios.The topic discovery experiments of multiple data source were conducted on public data sets,and a comparative analysis of topics was conducted.The experimental results showed that the confusion degree of IccLDA model is lower than LDA model and ccLDA model,indicating that IccLDA model has better modeling ability.Finally,further experimental verification was performed with the data sets of real-world scenarios.The result showed that the improved model not only has better modeling ability than the traditional models,but also can effectively discover public topics discussed by various data sources and local topics discussed by each data source,and is more suitable for topic discovery in multiple data source scenarios.

Key words: topic detection;topic model;LDA;multi-source;IccLDA

随着互联网的普及和信息技术的发展, 网络新闻、网络论坛、微信公众平台等新媒体已经深入人们的日常生活, 成为信息传播的重要渠道。网络舆情[1]是公众在网络媒体上公开表达的对某种社会问题或

收稿日期:2017 – 08 – 05

基金项目:国家科技支撑计划资助项目(2012BAH18B05); 国家自然科学基金资助项目(61272447); 四川省科技厅计划资助项目(16ZHSF0483)

作者简介:陈兴蜀(1968—), 女, 教授, 博士生导师, 博士. 研究方向: 云计算及大数据安全; 可信计算. E-mail: chenxsh@scu.edu.cn

* 通信联系人 E-mail: whzh.nc@scu.edu.cn

网络出版时间:2018 – 03 – 21 00 : 02 : 27　　**网络出版地址**: http://kns.cnki.net/kcms/detail/51.1773.TB.20180321.0002.001.html

社会现象具有一定影响力和倾向性的共同意见。网络舆情监管对于网络媒体的健康、有序发展具有重要意义，因此网络舆情监控关键技术的研究也成为了近年来的重要研究方向。新闻、论坛、微信公众平台每天会产生大量的文本数据，面对日益增长的海量互联网信息，通过多数据源话题检测技术，可以帮助相关人员了解不同媒体的共同关注点以及不同媒体是如何相互影响的，快速聚焦到公众讨论的热点。

热点话题检测技术属于话题检测与跟踪（TDT）的研究范畴，它主要的任务是从报道流中检测和识别未知的话题。

随着话题模型[2−3]的兴起，越来越多的学者将话题模型应用到话题检测当中，而Blei等[4]于2003年提出的潜在狄利克雷分配（LDA）话题模型，由于其优秀的话题建模能力，产生了很多改进模型。李湘东等[5]提出的加权LDA话题模型，通过对特征词加权，解决了LDA话题模型的话题分布向高频词倾斜的问题，提高了话题的表达能力。张越今等[6]对TDT领域应用较多的Single-pass算法进行改进，提出了一种基于相似哈希的增量型文本聚类算法，该算法相比于原Single-pass算法在聚类效率方面具有明显提升。相关话题模型（CTM）[7]和层次LDA模型[8]通过扩展模型参数，刻画了不同话题之间的相关性。Huang等[9]将LDA话题模型应用到微博短文本建模当中，然后利用Single Pass聚类算法进行话题检测。吴永辉等[10]将LDA与AP聚类算法结合，用于发现热点新闻话题。路荣等[11]利用LDA话题模型对Twitter数据建模，有效地解决了短文本集数据稀疏性的问题，然后利用*k*-means和层次聚类算法，可以有效的将微博聚集到不同的话题之下。高悦等[12]提出了一种基于狄利克雷过程混合模型的文本聚类算法，该算法基于非参数贝叶斯框架，可以将有限混合模型扩展成无限混合分量的混合模型。

针对多源话题的检测，Zhai等[13]首先提出比较性文本挖掘（CTM）的概念，CTM旨在发现可比文本集或者相似文本集之间话题的差异性，如话题在不同文化、时间、数据源所表现出来的差异性。然后，提出了基于PLSA的跨文本集混合模型（ccMix）。Xu等[14]首先利用LDA话题模型对各个文档集合进行建模，然后通过计算不同集合下话题之间的相似度，构建不同文档集合下话题之间的关系，该模型不足之处在于无法发现全局话题。Wang等[15]提出马尔科夫话题模型（MTM）实现多个语料库的话题发现，利用高斯随机场建模文档集合之间的话题关系，针对每个文档集合仍然采用LDA话题模型采样话题。谭文堂等[16]通过使用HDP信息检索模型和TF-IDF加权算法把话题划分为公共话题和文本集特有话题，解决了多个文本集中同时具有公共话题和文本集特有话题的问题。Miao[17]、Chen[18]等利用非参数贝叶斯的方法发现全局话题和每个数据源独有的话题，文本建模时不需要事先指定话题数。Deng等[19]利用文献中的文本内容、作者和会议信息这3种数据源，构建联合话题模型，在话题生成过程中，3种不同的数据源之间共享话题，实验表明联合不同数据源的话题建模效果优于单纯利用文本的传统建模方法。Paul等[20]针对文献[13]中话题模型存在手动设置参数过多和模型过拟合的问题，提出了基于LDA的跨文本集话题模型（ccLDA），ccLDA可以通过采样选择是通过采用全局的话题–单词分布生成单词，还是通过指定文本集合下的话题–单词分布生成单词，但该方法会强制全局话题和局部话题对齐，只适用于相似度很高的文本集，有很大的局限性。

相比已有的研究工作，作者针对ccLDA模型强制全局话题和局部话题对齐并只适用于文本集之间相似度很高的局限性，提出了改进的ccLDA模型（即Ic-cLDA模型）。IccLDA模型在采样过程中，先进行文档中词语的采样，判断其属于全局话题还是局部话题，再为该词语采样话题编号，克服了ccLDA模型中全局话题和局部话题必须对齐的缺点，进而避免了词语在全局话题和局部话题分散的问题，使其能够更加有效地找出全局话题和局部话题。

1 多数据源话题检测模型

1.1 ccLDA话题模型

Blei等基于LDA话题模型[4]，提出了跨数据源的LDA模型（cross-collection LDA，ccLDA）。在该模型下，每个话题关联两类话题–单词分布：一类是所有文档集合共享的话题–单词分布，即全局话题–单词分布；另一类是指定文档集合下的话题–单词分布，即局部话题–单词分布。该模型通过为文本中的每一个词语分别关联局部话题–单词分布和全局话题–单词分布，使得该模型能够捕获到不同文本集之间话题的差异性。与LDA话题模型相比，ccLDA模型在生成文档集合中的每一个单词时，首先为该单词采样一个话题编号，然后通过伯努利分布采样该话题是来源于全局话题–单词分布还是局部话题–单词分布，最后利用得到的话题–单词分布生成单词。

ccLDA话题模型通过一个隐话题将全局话题和局部话题关联起来使其共现，强制全局话题和局部话题对齐，话题数必须保持一致，使得该模型只适用

于文本集之间相似度很高的场景。然而, 在检测真实数据中的多数据源热点话题时, 每个文本集往往会有独特的话题, 这种情况下应用ccLDA话题模型就会导致很多不相关话题的对齐, 极大地降低了话题建模的效果。

1.2 改进的ccLDA模型(IccLDA)

针对ccLDA话题模型强制全局话题和局部话题对齐的问题, 作者提出了一种改进的ccLDA话题模型(IccLDA)。提出的IccLDA话题模型, 在采样文档中的每个词语时, 首先根据$\boldsymbol{\psi}_{s,m}$伯努利分布采样该词语是属于全局话题还是局部话题, 再采样该词语的话题编号。从本质上讲, ccLDA模型通过一个隐话题将全局话题和局部话题关联起来, 而IccLDA模型中的全局话题和局部话题不具有关联性, 这样可以有效地检测出多个数据源中全局话题和局部话题。

在描述具体算法之前, 首先给出算法中用到的符号, 如表1所示。

表 1　IccLDA算法中的符号

Tab.1　Symbol of IccLDA algorithm

符号	描述
C	数据源数
x	$x=c$表示全局话题, $x=s$表示局部话题
$\boldsymbol{\theta}_m^x$	文档m的全局(或局部)文档–话题分布
$\boldsymbol{\sigma}_k^{(s)}$	数据源s下话题k的话题–单词分布
$\boldsymbol{\varphi}_k$	全局话题k的话题–单词分布
γc、γs	全局话题和局部话题比例的超参数
$\boldsymbol{\psi}_{s,m}$	文档m中全局话题和局部话题比例的超参数
TC、TS	全局话题和每个数据源局部话题的数
x_i	该位置的词语是属于全局话题还是局部话题
C_m^x	文档m属于全局(或局部)话题的词语数

IccLDA话题模型的图模型表示如图1所示。

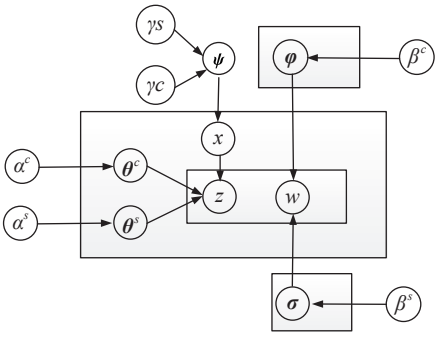

图 1　IccLDA图模型表示

Fig.1　IccLDA model

采样流程如下:

1)从以参数为β^c的Dirichlet分布中为每个全局话题采样话题–单词分布$\boldsymbol{\varphi}_k$, 即$\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\beta^c), k \in [1, TC]$;

2)针对每一个数据源s, 从以参数为β^s的Dirichlet分布中为数据源s中的每个局部话题采样话题–单词分布$\boldsymbol{\sigma}_k^{(s)}$, 即$\boldsymbol{\sigma}_k^{(s)} \sim \mathrm{Dir}(\beta^s), k \in [1, TS]$;

3)针对每一个数据源s中的每一篇文档m:

①从以参数为γs、γc的Beta分布为数据源s下的文档m采样伯努利分布$\boldsymbol{\psi}_{s,m}$, 即$\boldsymbol{\psi}_{s,m} \sim \mathrm{Beta}(\gamma s, \gamma c)$;

②从以参数为α^c的Dirichlet分布中为每个文档采样全局文档–话题分布$\boldsymbol{\theta}_m^c$, 即$\boldsymbol{\theta}_m^c \sim \mathrm{Dir}(\alpha^c), m \in [1, M]$;

③从以参数为α^s的Dirichlet分布中为每个文档采样局部文档–话题分布$\boldsymbol{\theta}_m^s$, 即$\boldsymbol{\theta}_m^s \sim \mathrm{Dir}(\alpha^s), m \in [1, M]$;

④针对文档m中每一个词语n: 首先, 采样$x_{m,n} \sim \boldsymbol{\psi}_{s,m}$; 随后, 采样一个话题标签$z_{m,n} \sim Mult(\boldsymbol{\theta}_m^{x_{m,n}})$; 最后, 如果$x_{m,n}=0$, 则根据$\boldsymbol{\varphi}_{z_{m,n}}$采样单词, 否则, 根据$\boldsymbol{\sigma}_{z_{m,n}}^{(s)}$采样单词$w_{m,n}$。

为了计算隐变量x和z, 作者使用吉布斯采样的方法进行近似推断, 每轮迭代按照上文中的采样流程重新分配话题编号, 话题编号的采样公式如式(1)、(2)所示:

$$p(x_i=c, z_i^c=k|w, z_{\neg i}) \propto \frac{C_{d,\neg i}^c+\gamma c}{N_m+\gamma c+\gamma s-1} \times \frac{n_{m,\neg i}^{(k)}+\alpha_k^c}{\sum_{k=1}^{TC}\left(n_{m,\neg i}^{(k)}+\alpha_k^c\right)} \times \frac{n_{k,\neg i}^{(t)}+\beta_t^c}{\sum_{t=1}^{V}\left(n_{k,\neg i}^{(t)}+\beta_k^c\right)} \tag{1}$$

$$p(x_i=s, z_i^s=k|w, z_{\neg i}) \propto \frac{C_{d,\neg i}^s+\gamma s}{N_m+\gamma c+\gamma s-1} \times \frac{n_{m,\neg i}^{(k,s)}+\alpha_k^s}{\sum_{k=1}^{TS}\left(n_{m,\neg i}^{(k,s)}+\alpha_k^s\right)} \times \frac{n_{k,\neg i}^{(t,s)}+\beta_t^s}{\sum_{t=1}^{V}\left(n_{k,\neg i}^{(t,s)}+\beta_k^s\right)} \tag{2}$$

由式(1)和(2)可知, 每轮迭代需要重新采样每个词语的话题分配, 采样收敛后, 计算文本–话题的后验分布以及话题–单词的后验分布, 计算公式如下:

$$\boldsymbol{\theta}_m^{k,c} = \frac{n_m^{(k)}+\alpha_k^c}{\sum_{k=1}^{TC}\left(n_m^{(k)}+\alpha_k^c\right)} \tag{3}$$

$$\boldsymbol{\theta}_m^{(k,s)} = \frac{n_{m,\neg i}^{(k,s)}+\alpha_k^s}{\sum_{k=1}^{TS}\left(n_{m,\neg i}^{(k,s)}+\alpha_k^s\right)} \tag{4}$$

$$\boldsymbol{\varphi}_t^k = \frac{n_k^{(t)}+\beta_t^c}{\sum_{t=1}^{V}\left(n_k^{(t)}+\beta_k^c\right)} \tag{5}$$

$$\boldsymbol{\sigma}_t^{(k,s)} = \frac{n_k^{(t,s)}+\beta_t^s}{\sum_{t=1}^{V}\left(n_k^{(t,s)}+\beta_k^s\right)} \tag{6}$$

2 实验及结果分析

为了验证IccLDA模型的有效性，使用Michael Paul的tourists数据集中的lonelyplanet部分（后面统称为tourists数据集）进行实验。tourists数据集是去过或者打算去英国、印度和新加坡的游客写的博客，博客内容包括游客在当地旅游的一些见闻以及准备去旅游的游客关心的一些问题。tourists数据集的具体信息如表2所示。

表 2 tourists数据集描述

Tab.2 Description of tourist dataset

国家	数量
India	1 432
Singapore	1 182
UK	1 580

针对tourists数据集，实验分别进行模型困惑度比较和话题比较性分析。

2.1 模型困惑度

模型困惑度用于衡量话题模型对于未观测数据的预测能力，困惑度越小，模型的预测能力越强，模型的推广能力越强。困惑度的定义如下：

$$pp(D_{\text{test}}) = \exp\left(-\frac{\sum_{d=1}^{M}\sum_{i=1}^{N_d} \text{lb}\, p\left(w_d^i\right)}{\sum_{d=1}^{M} N_d}\right) \tag{7}$$

式中，D_{test}为测试集，M为测试集的文档数，N_d为第d个文档的长度，$p(w_d^i)$为训练好的模型在第d个文档第i个词上生成的概率。

选取LDA模型、ccLDA模型和IccLDA模型在tourists数据集上做比较，训练集和测试集以9∶1的比例进行划分。其中：在LDA模型中，$\alpha = 50/K$，$\beta = 0.01$；在ccLDA模型中，$\alpha = 1.0$，$\beta = \delta = 0.01$，$\gamma_0 = \gamma_1 = 1.0$；在IccLDA模型中，$\alpha^s = \alpha^c = 1.0$，$\beta^s = \beta^c = 0.01$，$\gamma s = \gamma c = 1.0$。3个模型的迭代次数都为1 000。下面针对3个模型进行话题建模效果对比分析。

图2展示了话题数与困惑度之间的关系，实验显示IccLDA模型和ccLDA模型的困惑度低于LDA模型，说明这两个模型很好地利用了不同数据源的信息，以更高的概率将词语分配给其更可能出现的文档，而IccLDA模型的建模能力优于ccLDA，说明全局话题和每个数据源的局部话题对齐会影响文档的建模效果。上述3个话题模型随着话题数的增加，困惑度总体呈现一个下降的趋势，这是因为话题数量越多，在对文本进行降维时损失的信息就越少，降维结果就越符合文本集的特征。

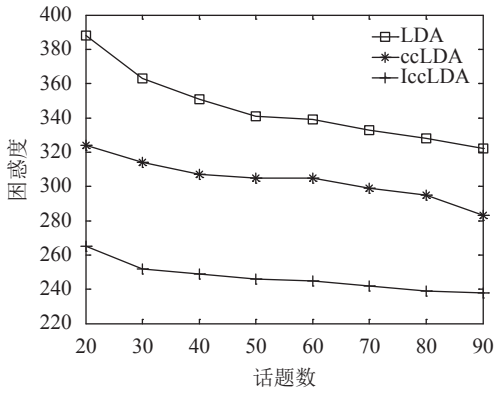

图 2 tourists数据集的模型困惑度对比

Fig.2 Models' perplexity contrast of tourists dataset

2.2 话题比较性分析

对LDA模型、ccLDA模型和IccLDA模型的话题进行详细的比较性分析。实验中，模型参数的设置和之前的一致，根据前面实验得到的经验值，这3个模型的话题数都设置为45。

1）LDA模型

在利用LDA模型对tourists数据集进行话题建模时，将印度、新加坡和英国这3个文本集看做同一文本集进行建模，得到的部分热点话题如表3所示。从表3中可以看出：话题39讨论了旅游中的住宿问题，话题5讨论了旅游中的交通问题，这些话题描述的都是一些旅游中遇到的常见问题和旅游中的见闻，都可能会出现在印度、新加坡和英国这3个文本集中，即全局话题；而话题25和话题29则分别讨论了关于新加坡和印度的问题，这些话题只会出现在新加坡或英国文本集中，即局部话题。由此可见，LDA建模并不能直接将全局话题和局部话题区分开。

表 3 LDA建模得到的话题

Tab.3 Topics of LDA model

话题编号	话题关键词
39	**hotel room hotels** stay stayed budget **rooms night place hostel** staying find price location inn **hostels** book **bed places** accommodation
19	**Singapore** mrt mind excuse cat food **India** sgd **zoo** night **orchard** road **Chinatown** Sentosa shopping city place area safari walk
9	visa UK passport immigration entry months apply country time enter work **travel** embassy return visas application stay **Australia**
25	**Singapore Malaysia** KL Asia taste island Indonesia Johor Thailand islands Kuala air ferry Hong Kong check travel China Lumpur
5	**airport** time **taxi** hours check luggage Changi transit **flight** hotel city hour morning leave immigration budget night **cab** minutes baggage arrive
29	**Delhi India Agra Varanasi** Jaipur time travel Rajasthan train Udaipur driver trip car **Taj** Jaisalmer weeks fort visit city pushkar

2）ccLDA模型

通过对ccLDA模型的分析得知，ccLDA话题模型只能适用于各个数据源的话题相似度很高的情况。印度、新加坡和英国这3个数据集都会涉及交通、金融、移民等问题，因此ccLDA话题模型针对这类公共话题有很好的建模效果，如表4所示，该话题在讨论关于交通的问题。但是，面对部分数据集在讨论的话题时，ccLDA的建模效果往往很差，因为它会将讨论不同事件的话题对齐，导致生成一些无意义的话题，如表5所示。

表 4　ccLDA建模得到的关于交通问题的话题

Tab.4　Topics about transportation of ccLDA model

话题类型	话题关键词（前15个）
全局话题	**flight airport** check **airlines airline** luggage book cost booking service **train** fare leave air arrive
India	Delhi India travel **car** driver time train book Mumbai trip **tickets trains fly flight** days
Singapore	**airport** Changi time **taxi** hours mrt transit city Singapore hotel budget excuse cat check hour
UK	Heathrow time **train ticket** London travel **tickets** hours **fly** buy **flights bus** frills express

表 5　ccLDA建模得到的非公共问题的话题

Tab.5　Local topics of ccLDA model

话题类型	话题关键词（前15个）
全局话题	places stay place visit days spend coast north night week end travel weather recommend find
India	goa beach Kerala time palolem cochin south days trip beaches Chennai India place Mysore varkala
Singapore	Singapore Malaysia island days time Sentosa travel visit beach bit live east city shopping water
UK	Edinburgh car time Scotland trip Skye drive route walk train castle driving miles walking lake

3）IccLDA模型

表6展示了IccLDA模型生成的部分全局话题。

表 6　IccLDA生成的全局话题

Tab.6　Global topics of IccLDA model

话题编号	话题关键词
8	**hotel room hotels** night **stay rooms place** price accommodation taxi book airport **places staying** booked area budget recommend **house stayed**
1	flight **airport** check **flights airlines** time **airline fly** Heathrow air **ticket** hours luggage travel book Ryanair London hour booked
14	**card money credit bank cash pay** phone give **atm account cards charge** fee rate check exchange number SIM currency buy
13	visa UK passport immigration entry months apply country enter time work return stay embassy application Australian leave visit tourist Australia
3	shops food tea buy shop carry bring find hand cafes shopping hours store bag cafe street breakfast stores supermarkets times

其中，话题8讨论了旅游目的地的住宿问题，话题1讨论了前往旅游目的地的交通问题，话题14讨论了在旅游地支付、取款的一些问题。结合表3和6可以看出，IccLDA模型和LDA模型生成的话题中包含一些共同话题，譬如关于“住宿”问题的讨论、关于“旅游交通”问题的讨论等等。

表7展示了IccLDA模型每个数据源生成的局部话题，三者主要在讨论当地的一些地点、景点和风俗等。由此可看出，IccLDA模型可以有效地将公共讨论的话题和每个文本集讨论的话题区分开。

表 7　IccLDA生成的每个数据源特有话题

Tab.7　Local topics of IccLDA model

India		Singapore		UK	
话题6	话题2	话题7	话题8	话题3	话题2
Kerala	**Delhi**	**Singapore**	food	**Dublin**	**London**
cochin	train	**Malaysia**	**Singapore**	city	bus
temple	time	Asia	place	**Ireland**	tube
Chennai	**Agra**	travel	eat	place	station
trip	travel	**Indonesia**	excuse	**castle**	day
visit	**Jaipur**	**island**	restaurants	history	time
places	**Varanasi**	**beach**	area	located	walk
India	**India**	ferry	bar	town	**Heathrow**
Madurai	**Rajasthan**	islands	restaurant	park	train
south	**Udaipur**	**Thailand**	sgd	island	**Birmingham**
Mysore	car	place	park	river	cross
fort	tour	trip	places	**museum**	buses
city	station	**tioman**	**seafood**	years	**museum**
Tamil	city	**bintan**	shops	gardens	guide
tour	**taj**	visit	stalls	street	hour

2.3　真实数据集验证分析

使用了从四川大学舆情监控系统中收集得到的数据，是利用主题爬虫以“四川大学”和“川大”为关键字采集得到的，包括新闻、论坛和微信公众平台3种数据源，随机选择了其中从2015年6月24日到2015年6月30日共6 580篇文档。各个数据源的数量描述如表8所示。

表 8　真实数据集描述

Tab.8　Description of the real dataset

数据源	数量
新闻	2 082
论坛	1 389
微信公众平台	3 109

图3展示了话题数与困惑度之间的关系，从图3中可以看出IccLDA模型的困惑度小于LDA模型和ccLDA模型，说明在真实数据集上IccLDA模型的建模能力仍然好于另外两个模型。

从图3中还可以看出，LDA模型、ccLDA模型和IccLDA模型随着话题数的增加，困惑度呈现一个下降的趋势，这是因为话题数越多，在对文本进行降维时损失的信息就越少，降维结果就越符合文本集的特征。但是如果话题数过多，文本集的话题解释性就会变差，通过实验观察，在真实数据集上话题数取50效果最佳。

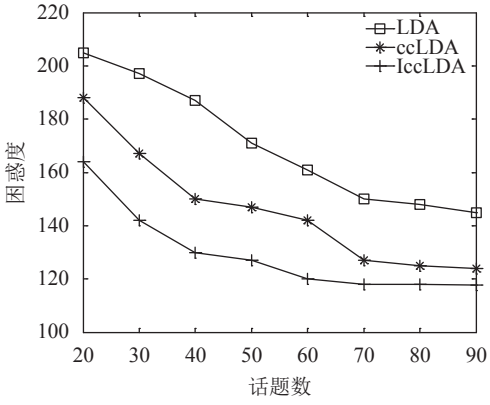

图 3 真实数据集的模型困惑度对比

Fig.3 Models' perplexity contrast of real dataset

使用IccLDA模型对上述真实数据集进行了话题检测，模型的参数设置和上文保持一致，迭代次数为1 000，话题数设置为50。

表9和10是IccLDA生成的全局热点话题和局部热点话题。

表 9 IccLDA生成的全局话题

Tab.9 Global topics of IccLDA model

话题	词语
8	四川 大学 **泸州** 医科 **更名** 华西 **医学院** 教育部 学校 学生 行政 改名 反对 **校名** 四川省 复议 医学 同意 简称 教育
16	**谢和平** 校长 四川 **毕业** **毕业生** 精神 社会 希望 大学 **典礼** 精力 学生 功名 权威 工作 追逐 **学子** 生活 同学 人生
14	专业 志愿 招生 录取 考生 四川 大学 学校 填报 电子 高考 分数 本科 成绩 学院 学生 计划 华西 家长 原则
15	大学 中国 专业 学科 工程 实力 北京 全国 上海 质量 华东 优势 华中 华南 名校 排行榜 教学 南京 学校 校友
13	荷花 大学 校园 荷塘 望江 四川 苏州 中国 建筑 成都 拍摄 旅游 照片 地址 相机 风景 阅读 厦门 规模 江南

从表9中可以看到“泸州医学院更名”和“川大毕业典礼”等新闻、论坛、微信均有讨论的全局热点事件。

表10中展示了新闻、论坛、微信中讨论的局部热点事件，如新闻中的“天府新区的创新创业园区建设”事件，论坛中的“抗战文化宣传”事件，微信中的“院校招生志愿填报”事件等。

表 10 IccLDA生成的局部话题

Tab.10 Local topics of IccLDA model

新闻		论坛		微信	
话题4	话题6	话题2	话题3	话题9	话题1
网络	成都	四川	川剧	志愿	本科
信息	**创新**	大学	抗战	天津	**志愿**
中国	**创业**	中国	活动	大学	高考
企业	**天府**	重庆	四川	车站	录取
公司	**新区**	**抗战**	精神	四川	**填报**
互联网	**科技**	**历史**	调研	专业	**分数线**
行业	大学	**川剧**	宣传	北大	**考生**
市场	**产业**	高校	纪念	座位	理科
李涛	国家	**博物馆**	成都	纸条	**招生**
国家	中心	学院	采访	平行	文科
技术	四川	学校	人民	同学	**成绩**
服务	服务	**文化**	文化	上车	**院校**
数据	示范区	成都	传统	汽车	专科
平台	自主	资料	艺术	停车场	提前
科技	技术	**巴蜀**	方式	排队	考试

3 结 论

随着互联网的普及和信息技术的发展，网络新闻、网络论坛和微信公众平台等新媒体吸引了众多网民的参与，成为信息传播的重要渠道。新闻、论坛和微信公众平台每天会产生大量的文本数据，面对日益增长的互联网信息，作者在单一数据源话题检测的基础上，研究了多数据源话题检测技术。

首先，讨论了ccLDA话题模型的不足之处，该模型只适用于各个数据源话题相似度很高的情况，且因为全局话题和各数据源的局部话题对齐会导致词语稀疏的问题。针对以上缺陷，作者提出了IccLDA模型。该模型增加了对词语进行判断属于全局话题还是局部话题的过程；然后，进行采样，避免了全局话题和局部话题的强制对齐等问题；最后，分别在公开数据集和真实数据集上进行了实验，验证了模型的有效性，证明了IccLDA模型不仅能发现多个数据源讨论的全局话题，还能发现各个数据源分别讨论的局部话题。

但是，所提出的IccLDA模型中，在采样每个词语之前，会利用伯努利分布判断该位置属于全局话题还是局部话题，没有充分利用文本集合的先验知识。因此，在下一步研究中，将尝试引入更加智能的词语采样策略，以及引入新的数据源研究更全面的话题检测方法。

参考文献:

[1] Xu Xiaori.The research on emergency treatment of internet public opinion events[J].Journal of North China Electric Power University (Social Science Edition),2007(1):89–93.[徐晓日.网络舆情事件的应急处理研究[J].华北电力大学学报(社会科学版),2007(1):89–93.]

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

[3] Xu Ge,Wang Houfeng.The development of subject model in natural language processing[J].Journal of Computers,2011,38(08):1423–1436.[徐戈,王厚峰.自然语言处理中主题模型的发展[J].计算机学报,2011,38(08):1423–1436.]

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

[5] Li Xiangdong,Ba Zhiyao,Huang Li.Based on the weighted implicit Dirichlet distribution model of news topic mining method[J].Computer Applications,2014,34(5):1354–1359.[李湘东,巴志超,黄莉.基于加权隐含狄利克雷分配模型的新闻话题挖掘方法[J].计算机应用,2014,34(5):1354–1359.]

[6] Zhang Yuejin,Ding ding.A study on incremental text clustering in sensitive topic detection[J].Netinfo Security,2015(9):170–174.[张越今,丁丁.敏感话题发现中的增量型文本聚类模型[J].信息网络安全,2015(9):170–174.]

[7] Blei D M,Lafferty J D.A correlated topic model of science[J].The Annals of Applied Statistics,2007,1(1):17–35.

[8] Blei D,Griffiths D,Jordan D,et al.Hierarchical topic models and the nested Chinese restaurant process[C]//Proceedings of the 7th Annual Conference on Neural Information Processing Systems.Vancouver:MIT Press,2003:17–24.

[9] Huang B,Yang Y,Mahmood A,et al.Microblog topic detection based on LDA model and single-pass clustering[C]//Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing.Chengdu:Springer,2012:166–171.

[10] Wu Yonghui,Wang Xiaolong,Ding Yuxin,et al.Topic-based adaptive,online hotspot discovery method and news recommendation system[J].Journal of Electronics,2010,8(11):32620–2624.[吴永辉,王晓龙,丁宇新,等.基于主题的自适应、在线网络热点发现方法及新闻推荐系统[J].电子学报,2010,8(11):32620–2624.]

[11] Lu Rong,Xiang Liang,Liu Mingrong,et al.Discovery of news topics based on hidden subject analysis and text clustering in microblogging[J].Pattern Recognition and Artificial Intelligence,2012,25(3):382–387.[路荣,项亮,刘明荣,等.基于隐主题分析和文本聚类的微博客中新闻话题的发现[J].模式识别与人工智能,2012,25(3):382–387.]

[12] Gao Yue,Wang Wenxian,Yang Shuxian.A document clustering algorithm based on Dirichlet process mixture model[J].Netinfo Security,2015(11):60–65.[高悦,王文贤,杨淑贤.一种基于狄利克雷过程混合模型的文本聚类算法[J].信息网络安全,2015(11):60–65.]

[13] Zhai C X,Velivelli A,Yu B.A cross-collection mixture model for comparative text mining[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.Seattle:DBLP,2004:743–748.

[14] Xu Shuo,Zhu Lijun,Qiao Xiaodong.Topic linkages between papers and patents[C]//Proceedings of the 4th International Conference on Advanced Science and Technology.Shanghai:Science and Engineering Research Support Society,2012:176–183.

[15] Wang C,Thiesson B,Meek C,et al.Markov topic models[J].Journal of Machine Learning Research,2009,5(2):583–590.

[16] Tan Wentang,Wang Zhenwen,Yin Fengjing,et al.A partial comparative LDA model for multi-text set[J].Computer Research and Development,2013,50(9):1943–1953.[谭文堂,王桢文,殷风景,等.一种面向多文本集的部分比较性LDA模型[J].计算机研究与发展,2013,50(9):1943–1953.]

[17] Miao Y,Li C,Ding Q,et al.Research on mining common concern via infinite topic modelling [C]//Proceedings of the the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology.Macau:IEEE,2012:180–184.

[18] Chen C,Buntine W,Ding N,et al.Differential topic models[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(2):230–242.

[19] Deng H,Zhao B,Han J.Collective topic modeling for heterogeneous networks[C]//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval.Beijing:ACM,2011:1109–1110.

[20] Paul M.Cross-collection topic models:Automatically comparing and contrasting text [D].Urbana:University of Illinois at Urbana-Champaign,2009.

（编辑　赵　婧）

引用格式: Chen Xingshu,Ma Chenxi,Wang Wenxian,et al.Multi-source topic detection analysis based on improved ccLDA model[J].Advanced Engineering Sciences,2018,50(2):141–147.[陈兴蜀,马晨曦,王文贤,等.基于改进的ccLDA多数据源热点话题检测模型[J].工程科学与技术,2018,50(2):141–147.]