融入事件知识的主题表示方法

孙 锐[1),3)] 郭 晟[1)] 姬东鸿[1),2)]

[1)](武汉大学计算机学院 武汉 430072)

[2)](武汉大学软件工程国家重点实验室 武汉 430072)

[3)](乐山师范学院计算机科学学院 四川 乐山 614000)

摘 要 主题模型已被广泛用于发现文档潜在主题.已有方法多采用词或短语来表示主题,然而这些方法生成的主题缺乏深层次的语义信息,可解释性比较差.文中提出使用结构化的事件来表示主题.一方面,事件包含比词或短语更丰富的语义;另一方面,一组相关的事件能更合理地解释并区分不同的主题.为解决事件作为基本单元所带来的稀疏性问题,该文在Biterm Topic Model(BTM)的基础上提出两种主题模型,采用两种不同的方式将事件的语义知识融入到主题生成过程中.其中,第1种模型利用Generalized Pólya Urn(GPU)模型天然的聚类效果加大语义相近的事件分配到同一主题的概率,而第2种模型则通过为每个biterm引入指示变量,合理地利用语义知识有效地解决同一个biterm中两个事件的主题分配问题.该文不仅从主题凝聚度和KL散度两个指标直接对主题模型进行评估,还通过将主题表示结果引入到文本分类任务中对模型进行了外部评估.实验结果表明文中提出的模型从共现和语义两个层面有效地解决了事件稀疏性问题.与基于词或短语的主题表示相比,事件结构所包含的语义信息提高了主题生成质量,使主题表示具有更强的可读性和主题判别性.

关键词 事件;主题模型;主题表示;事件知识;自然语言处理;社交网络;社会媒体

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2017.00791

Topic Representation Integrated with Event Knowledge

SUN Rui[1),3)] GUO Sheng[1)] JI Dong-Hong[1),2)]

[1)](*Computer School, Wuhan University, Wuhan* 430072)

[2)](*State Key Laboratory of Software Engineering, Wuhan University, Wuhan* 430072)

[3)](*School of Computer Science, Leshan Normal University, Leshan, Sichuan* 614000)

Abstract Topic model has been widely used to discover the latent topic of text. Most previous methods exploited words or phrases for topic representation. However, this form of topic representation has a poor interpretability, due to the lack of deep semantic information. This paper proposes to exploit structured events for topic representation. On one hand, events have more abundant semantic information than words or phrases; on the other hand, a set of events are able to interpret and distinguish different topics intuitively. However, the structured events, as basic units of document, add more difficulties to the topic sampling because of the sparseness. To address the problem, we propose two topic models based on Biterm Topic Model. Event semantic knowledge is incorporated into these models using two different ways. The first model exploits the natural clustering performance of Generalized Pólya Urn model to increase the probability of assigning same topic to similar events. Differently, the second model introduces an indicator variable for each biterm, and exploits event semantic information to solve the topic assignment of the events in one biterm more reasonably. We not only directly evaluate the topic models based on

收稿日期:2016-05-02;在线出版日期:2016-09-08.本课题得到国家自然科学基金(61133012,61373108,61373056)、国家社科重大招标计划项目(11&ZD189)资助.**孙 锐**,男,1977年生,博士研究生,主要研究方向为自然语言处理和机器学习.E-mail:ruisun@whu.edu.cn.**郭 晟**,男,1992年生,硕士研究生,主要研究方向为自然语言处理.**姬东鸿**(通信作者),男,1967年生,教授,博士生导师,主要研究领域为自然语言处理、信息检索.E-mail:dhji@whu.edu.cn.

two metrics, namely topic coherence and KL-divergence, but also conduct the external evaluation by carrying out text classification task based on the results of topic representation. The experimental results demonstrate our topic models effectively diminish the sparseness from two perspectives: event co-occurrence and semantic relatedness. Compared to the topic representation based on words, the semantic information of event effectively promotes the topic quality and improves the interpretability and topic discrimination of topic representation.

Keywords event; topic model; topic representation; event knowledge; natural language processing; social networks; social media

1 引 言

主题模型及其扩展模型、Probabilistic Latent Semantic Analysis (pLSA)[1] 和 Latent Dirichlet Allocation(LDA)[2] 等,已被广泛用于发现文本主题.这些模型大多基于"词条同现"并以词为基本单元[3].生成的每个主题被看成是在词上的分布,每个主题下高概率的词条被用于表示主题.大量研究结果表明这类模型有着较强的实践意义.但不可避免地,最后得到的主题词为孤立的语义单元,丢失了原始文档中词间的语义关系.这种基于词的表示方法主要存在以下缺点:

(1)一些领域性较强的词难以解释,非领域专家很难理解这些主题词的具体含义;

(2)部分主题词汇难以单独使用,需要在特定的语义环境下才能有合理的解释;

(3)部分短语被切分为多个个体词后会引起额外的重现,使得生成的主题中可能出现一些无关词汇从而导致主题可读性下降.

总的来讲,基于词的主题表示具有不可解释、可读性差等问题.为解决此问题,一些研究工作开始关注如何在主题模型中引入更高阶的语义单元.文献[4]在主题模型的词条中尝试加入短语.文献[5]和文献[6]则以事件为出发点,将 LDA 模型中的词替换为形如(subject、predicate)或(predicate1、predicate2)的事件或者是单个的动词.这些工作在一定程度上解决了基于词的主题表示所存在的语义缺失问题,但并未关注如何解决高阶语义单元共现稀疏的问题.

事件作为一种结构化的表示,拥有比词汇更丰富的语义.使用事件作为篇章学习的基本单元,已广泛用于一些 NLP 任务并取得良好的效果[7-9].因而,用事件作为主题描述的基本单元对于提升主题表示的可解释性也有着深远的意义.然而,以基于事件的主题模型所面临的主要挑战之一就是稀疏性问题.

与基于词的主题模型相似,每一篇文档被视为事件的集合,因此文档中的元素将变得更为稀疏.这种事件共现的稀疏性必然会给主题模型的学习带来一定的难度.直观来看,稀疏事件集合所组成的文档可以看成一篇短文本.基于 biterm 的主题模型(Biterm Topic Model,BTM)[10]能在一定程度上针对共现对 biterm 建模缓解词条共现的稀疏性.受此工作的启发,本文亦针对事件 biterm 结构进行建模.与传统 BTM 相比,本文提出的模型主要有以下区别:

(1)应用于短文本的 BTM 假设文档仅有一个主题,因而模糊了文档的概念.本文中由事件集合组成的"短文本"实际上是长文本压缩而成,因而我们在模型中保留了文档层.

(2)BTM 中假设同一个 biterm 中的两项均为同一主题.本文提出的第 2 种模型通过引入新的隐式变量允许 biterm 中的两个"词条"有不同的主题.

(3)本文通过引入事件相关性的外部先验知识,进一步从语义层面上来解决稀疏性问题.

基于事件 biterm 结构,本文提出两种在主题模型中融入事件知识的建模方法.第 1 种方法(Event_BTM+GPU)基于 Generalized Pólya Urn(GPU)[11] 模型,利用相关性事件采样来缓解稀疏性.这种方案简单直接,有效地利用了长尾事件.事件 biterm 中两个事件距离不一定连续,主题并不一定相同.因而,与 BTM 中"一个 biterm 内词条主题相同"的假设不符.针对此问题,本文提出第 2 种方法(Event_nBTM),在贝叶斯网络中引入新的隐式变量以确定 biterm 中两个"词条"分配主题的方式.与现有的相关研究工作相比,本文的主要贡献在于:

(1)提出采用三元组事件作为主题表示的基本单元.同基于词的表示方法相比,结构化的事件包含了词间的语义联系,因而拥有更细粒度的语义.中文语料上的实验结果表明,基于事件的主题表示包含了较强的语义信息,可读性强.

(2)为解决高阶语义单元所带来的稀疏性问题，本文提出两种融入事件知识的主题模型，从共现和语义相关两个方面有效地降低事件稀疏性的影响.对比实验结果表明，这两种模型都取得了较好的主题聚类效果.

(3)目前学术界尚未存在公开的事件相关的知识库，本文利用基于背景语料习得的分布式表示来定义事件相关性的先验知识.实验结果表明，这种事件知识的引入有效地提升了主题质量，同时也缓解了开放事件抽取性能所带来的负面影响.

2　相关工作

2.1　基于词的主题模型

多数主题模型都是通过词在文档中的共现信息来生成语义相关的主题[12].针对基于词汇的模型所生成的主题语义不够明确的问题，已有一些研究工作正在关注如何使主题中词项的语义更为明确从而提升主题生成的质量.

文献[13]基于相关上下文提出针对通用词和特殊词的两种分布，并使用上下文潜在语义分析模型用以挖掘上下文信息.基于此工作，文献[14]在情感分析任务中对模型进行扩展，提出主题情感混合模型(TSM)，在主题生成中进一步考虑了对正面、中性和负面等情感主题词的采样.为从商品评论数据中抽取能表示商品特征的词汇，文献[15]将特征词汇分为全局和局部两类，构建多粒度主题模型MG-LDA.文献[16]在特征词的基础上进一步引入特征修饰观点词，将词汇分为背景词、通用特征词、特殊特征词、通用观点词和特殊观点词等类型，并采用最大熵模型进行词类判别，建立联合特征和观点抽取的MaxEnt-LDA模型.以上这些工作仍以词为主题模型的基本单元，关注的重点是如何使得主题词的语义更明确，以得到更高的主题凝聚度.

另外有一些研究利用文本的词间关系来构建主题模型.文献[17]认为每个词的生成不仅与主题有关还与其前一个词有关，据此结合二元语言模型提出二元主题模型.为解决微博、推特等短文本中词汇共现稀少的问题，文献[10]提出BTM模型.该模型首先抽取一个短文本窗口内出现的所有词的共现词对(biterm)，然后基于biterm进行主题建模，并假定每个词对中的两个词有相同的主题.

2.2　基于高阶语义单元的主题模型

改善主题语义的另一个方向是关注比词汇更高阶的语义单元.一般都在主题模型中将传统的词分布替换为高阶语义单元的分布.部分研究工作在主题模型中加入短语或事件信息.

文献[18]提出n元主题模型(Topical N-Gram Model，TNG)，实现了词和短语在主题生成中的统一.文献[4]在主题模型的“词条”中加入了预先从语料中抽取出的名词短语.他们认为将主题看成是在单个词的分布上有着明显的缺陷，将两个词构成的短语视为等价于两个独立的词，会导致不正确的同现统计，因此多词短语应被视为一个“词条”.

文献[5]在主题模型中引入事件的概念.其基本思想是在LDA模型中只关注事件触发词(动词)，因此这种方法仍存在传统主题表示中单个词汇的语义与主题语义信息脱离的问题.类似地，文献[6]也基于LDA模型提出将文档看成事件的集合，其事件的定义为二元组，因此“词条”的表示形式为(subject、predicate)或(predicate1、predicate2).与文献[5]相比，这个工作关注了更丰富的语义结构，在日文语料中取得了良好的效果.不同于以上的工作，本文的事件形式是语义更丰富的三元组，并且特别地关注如何解决更复杂基本单元所带来的稀疏性问题.

2.3　融入先验知识的主题模型

近年来的一些研究[19-21]提出在主题模型中加入时间、社交网络数据和引用信息等先验知识并产生更连贯的主题.这些模型大多面向微博或科技文献等领域，因此这些知识存在领域依赖并且需要假设都是正确的.此类方法的一个主要缺陷就是要求用户对领域非常熟悉并能提供适合主题的知识.相对于领域知识而言，词汇知识更具有一般性.

文献[22]提出DF-LDA模型，采用引入两种先验知识，即Must-links和Cannot-links.其中，Must-links表示两个相连接的词在所有主题下的先验概率相近，而Cannot-links表示两个相连接的词不应该属于相同主题.文献[23]借助马尔可夫随机场(Markov Random Field，MRF)，使用一阶谓词逻辑表达词汇知识，建立了融入MRF的主题模型.由于MRF中的势函数定义灵活，便于表达各类知识，因此，MRF开始成为主题模型中融入先验知识的一种主要方法.文献[24]借助词汇相关性，在主题层利用MRF在相似词的主题变量间建立连接.文献[25]则定义了能同时考虑Must-links和Cannot-links知识的势函数，构建SC-LDA模型，在增大为Must-links连接词分配同一主题的概率的同时，也降低为Cannot-links连接词分配相同主题的概率.

另外有一些研究则关注在主题推理过程中加入先验知识.采用 Generalized Pólya Urn(GPU)模型来融入先验知识是一种直接有效的方法.其基本思想是:以瓮中取球模拟采样过程,简单的 Pólya Urn 模型在取出一个颜色为 w 的球后,放回并额外的放入一个颜色为 w 的球;而 Generalized Pólya Urn 模型则是在取出一个颜色为 w 的球后,放回并额外放入若干个颜色为 v 的球.这些额外放入的不同颜色的球的个数由其颜色 v 和 w 的相关度来决定.因此 GPU 可以保证后续采样中不仅能以更高的概率再次取到相同颜色的球,同时也能以较高的概率取到其他相似颜色的球.文献[26]在 LDA 中引入 GPU 模型生成连贯的主题,但并未使用任何外部资源或知识.文献[27]提出 MDK_LDA 模型,在采样中借助 GPU 模型融入同义词知识,加大同义词分配到相同主题的概率.文献[28]则在 MDK_LDA 的基础上进行了改进,利用通用知识消除了领域依赖并能显式地处理错误知识.文献[29]在 GPU 模型中考虑到多个瓮间的交互,利用领域知识实现主题数目的自动调整,提出的 MC_LDA 模型在观点词抽取任务中得到了有效的验证.

3 方 法

本文提出的方法首先需要将文档处理成事件集合.因此,我们从每篇文档中提取事件,并将每个事件定义为文本中的"词条".本节首先讨论了事件抽取和事件知识获取方法,然后详细介绍几个基于事件的主题表示模型.表 1 给出了本文模型中使用的基本符号及其含义.

表 1 基本符号及含义

符号	含义
α	超参,文档主题分布先验
β	超参,事件主题分布先验
φ	隐变量,事件主题分布
θ	隐变量,文档主题分布
γ	由事件知识生成的伯努利分布先验
V 或 v	词典大小(小写表示序号)
K 或 k	主题个数(小写表示序号)
M 或 m	文档总数(小写表示序号)
Z 或 z	主题分配隐变量(小写表示当前主题分配)
B 或 b	共现对 biterm 集合(小写表示当前 biterm)
L 或 l	隐变量,取值 0 或 1,用以确定当前 biterm 分配主题方式
E 或 e	事件集合(小写表示当前事件)
i,j	序号
t	主题编号
n	计数器

3.1 事件抽取与事件知识

3.1.1 事件抽取

事件起源于认知科学.认知科学家认为,以"事件"为单位来体验和认识世界符合人们正常的认知规律.在 ACE 评测会议中,"事件"被描述为一个动作的发生或状态的变化.但目前学术界对"事件"仍没有统一的定义,不同领域对"事件"的理解不同.综合近年来事件抽取和事件应用的研究中对事件的定义,事件大多被表示为谓词+论元结构[30-33].本文研究的"事件"为原子事件,我们希望能抽取出每个语句中的多个事件,事件所对应的谓词+论元结构表示为(Subject,Predicate,Object).本文采用开放领域的事件抽取方法.不同于传统的事件抽取任务,开放事件抽取没有预定义的事件类型.这里我们沿用文献[9]和文献[31]的方法抽取事件.给定文本,首先使用 Stanford 依存分析工具获得每条语句的依存结构;然后主要利用两种依存关系(nsubj 和 dobj)来抽取事件.如果两个 nsubj 和 dobj 关系拥有相同的谓词,则可合并为一个事件,并表示成元组(Subject,Predicate,Object).例如,给定语句"中国气象局启动台风二级应急响应",可以从其依存结果中获取两个依存对:"nsubj(启动-3,气象局-2)","dobj(启动-3,响应-8)".这两个依存对可合并为事件"(气象局,启动,响应)".对于依存结果中的部分未合并关系,仍保留为二元组事件.

3.1.2 事件知识

与词汇不同,目前并不存在公开的事件相关的知识库.本文采用事件相似度来表示事件的关联性,并以此作为事件先验知识加入到主题模型.

文献[34]引入三元组作为背景知识的一种表现形式并成功用于从外部资源中获取相关知识.其知识向量的定义来源于三元组内各成分词的主题分布.近年来,大量的研究工作表明分布式向量在很多 NLP(Natural Language Processing)应用中都显示了良好的性能.本文采用分布式向量来表示每个事件.每个三元事件的向量采用文献[35]提出的表示方案:

$$\overrightarrow{Sub,Verb,Obj} = \overrightarrow{Verb} \cdot (\overrightarrow{Sub} \otimes \overrightarrow{Obj}) \quad (1)$$

其中:$\otimes$表示克罗内克积运算;$\cdot$表示点乘运算.即每个三元事件的组合语义需要事件谓词向量去点乘事件主语和宾语向量的克罗内克积.对抽取的二元事件,则直接采用谓词向量和论元向量点乘的结果来表示.

本文中词的分布式向量通过基于 CBOW 模型的 word2vec[①] 工具在中文 Giga 语料上训练得到. 基于这种通过大规模语料训练得到的分布式向量可以习得词间的相关性，因而，基于语义组合的事件表示也是有意义的. 本文的事件相似度计算采用事件向量的欧式距离.

3.2　基线模型

3.2.1　Event_LDA 模型

LDA 是一种生成模型，文档的主题分布可通过每个主题出现在文档中的概率来得到，而每个主题均被表示成“词条”的多项式分布.

与文献[5]和文献[6]的工作类似，Event_LDA 模型以事件作为“词条”，为每个事件分配主题，因此每个主题被表示成事件的多项式分布.

3.2.2　Event_BTM 模型

传统 Biterm Topic Model(BTM)模型假设文档只有一个确定的主题，因而并没有生成文档主题. 该模型假定 biterm 集合中的每个共现对的主题都来自于同一个分布. 为缓解数据稀疏对使用 Event_LDA 模型进行主题生成的影响，我们借鉴 BTM 中的 biterm 结构进行主题建模，构建了 Event_BTM 模型. 图 1 所示为 Event_BTM 所对应的概率图，其中所有文档中包含的事件 biterm 构成的随机变量 B 为可观察变量，所有事件 biterm 对应的主题变量 Z 和主题事件分布参数变量 φ 以及各文档的主题分布参数 θ 为隐变量.

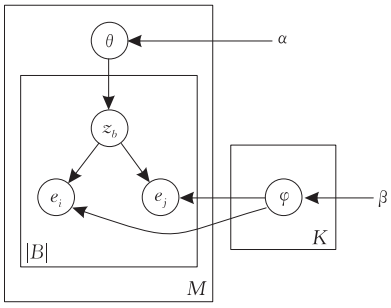

图 1　Event_BTM 模型

由此，该模型的生成过程可描述如下：

(1) 对于每个主题 $k\in\{1,2,\cdots,K\}$，生成主题“词条”的多项分布参数 $\varphi_k\sim Dir(\beta)$；

(2) 对于每篇文档 $m\in\{1,2,\cdots,M\}$：

1) 生成文档(biterm 集合)主题多项分布参数 $\theta_m\sim Dir(\alpha)$；

2) 对于文档 m 中的每个“词条”共现对 $b(e_i,e_j)$：

① 采样生成主题 $z_b\sim Mult(\theta_m)$；

② 采样生成“词条”$e_i\sim Mult(\varphi_{z_b})$，$e_j\sim Mult(\varphi_{z_b})$.

其联合概率可如下定义：

$$p(B,Z,\theta,\varphi|\alpha,\beta)=\prod_{k=1}^{K}p(\varphi_k|\beta)\times\prod_{m=1}^{M}p(\theta_m|\alpha)\times\prod_{b(e_i,e_j)\in B}p(z_b|\theta)p(e_i|\varphi_{z_b})p(e_j|\varphi_{z_b}) \quad (2)$$

3.3　融入事件知识的 Event_BTM+GPU 模型

3.3.1　Event_BTM 模型的采样和推理

BTM 模型大多采用 Collapsed Gibbs Sampling 的方法来辅助完成统计推理. 该方法是马尔可夫链蒙特卡洛方法的一种，通过使用随机样本来模拟概率模型. 按照吉布斯采样的步骤，BTM 模型以 $p(Z|B;\alpha,\beta)$ 为目标分布，每次对高维随机变量的一个分量在其他变量的条件下采样，即每次对 biterm 集合组成的高维变量 B 中某个事件共现对 b 的主题变量 z_b，按照概率 $p(Z_b|Z_{-b},B;\alpha,\beta)$ 进行采样，取经过一定迭代次数马尔可夫链收敛后 Z 的样本作为每个 biterm 的主题分配. 该后验概率正比于条件先验概率 $p(Z_b|Z_{-b};\alpha)$ 与条件似然概率 $p(b|z_b,Z_{-b},B_{-b};\beta)$ 的乘积，因此其计算公式如式(3)：

$$\begin{aligned}p(z_b=t|Z_{-b},B;\alpha,\beta)&\propto p(b|z_b,Z_{-b},B_{-b};\beta)\times p(z_b|Z_{-b};\alpha)\\&=\int p(b|z_b,\varphi)p(\varphi|Z_{-b},B_{-b},\beta)\mathrm{d}\varphi\times\int p(z_b|\theta)p(\theta|Z_{-b},\alpha)\mathrm{d}\theta\\&=\frac{\Gamma\left(\sum_{v=1}^{V}n_{t,v}^{-b}+\beta\right)\prod_{v=1}^{V}\Gamma(n_{t,v}+\beta)}{\prod_{v=1}^{V}\Gamma(n_{t,v}^{-b}+\beta)\Gamma\left(\sum_{v=1}^{V}n_{t,v}+\beta\right)}\times\frac{\Gamma\left(\sum_{k=1}^{K}n_k^{-b}+\alpha\right)\prod_{k=1}^{K}\Gamma(n_k+\alpha)}{\prod_{k=1}^{K}\Gamma(n_k^{-b}+\alpha)\Gamma\left(\sum_{k=1}^{K}n_k+\alpha\right)}\\&=\frac{n_{t,e_i}^{-b}+\beta}{\sum_{v=1}^{V}(n_{t,v}^{-b}+\beta)}\times\frac{n_{t,e_j}^{-b}+\beta}{\sum_{v=1}^{V}(n_{t,v}^{-b}+\beta)+1}\times\frac{n_t^{-b}+\alpha}{\sum_{k=1}^{K}(n_k^{-b}+\alpha)}\end{aligned} \quad (3)$$

其中：n_{t,e_i} 和 n_{t,e_j} 分别表示 e_i 和 e_j 对应“词条”分配到主题 t 的频次；n_t 表示 biterm 集合中的共现对分配到主题 t 的频次；$-b$ 表示集合中去除当前共现对 b 后的剩余项.

① https://code.google.com/p/word2vec/

3.3.2 Event_BTM+GPU 模型

式(3)中的因子对应了 Pólya Urn 模型产生的分布,从瓮中取球的物理过程可以看出,GPU 具有“马太效应”.直观来看,这种基于计数的方式会使得主题分布向高频“词条”倾斜.这种天然的聚类方法使得同一主题中包含更多语义相关的“词条”.因此,加大语义相近的“词条”分配到同一主题的概率,能在一定程度下更多的利用主题中的长尾词汇.同时,也能部分消除基本模型中“词条”共现所存在的稀疏性问题.

同文献[27]类似,借助 GPU 模型,可以实现在采样过程中不仅能以更高的概率取到对应的“词条”,同时也能以更高的概率取到其他相关的“词条”.由此可见,采样的效果实际上是由先验知识矩阵 $\boldsymbol{A}$ 来决定.相应地,式(3)中的似然概率 $p(b|z_b=k,Z_{-b},B_{-b};\beta)$ 可更改为如下形式:

$$\frac{\sum_{v=1}^{V} n_{k,v}^{-i}\boldsymbol{A}_{e_i,v}+\beta}{\sum_{v'=1}^{V}\left(\sum_{v=1}^{V} n_{k,v}^{-i}\boldsymbol{A}_{v',v}+\beta\right)}\times\frac{\sum_{v=1}^{V} n_{k,v}^{-j}\boldsymbol{A}_{e_j,v}+\beta}{\sum_{v'=1}^{V}\left(\sum_{v=1}^{V} n_{k,v}^{-j}\boldsymbol{A}_{v',v}+\beta\right)} \tag{4}$$

为了增大语义相近的元素分配到相同主题的概率,我们可通过事件相似度对矩阵 $\boldsymbol{A}$ 赋值如下:

$$\boldsymbol{A}_{e_i,e_j}=\begin{cases} sim(e_i,e_j), & sim(e_i,e_j)>\sigma \\ 0, & \text{其他} \end{cases} \tag{5}$$

其中:$sim(e_i,e_j)$ 为 e_i 和 e_j 两事件的相似度;σ 为指定的相似度阈值.

在 Event_BTM 模型中利用 GPU 模型融入事件语义知识(新模型被命名为 Event_BTM+GPU).我们可以通过设置阈值 σ 来调整采样的时间复杂度.由相似的对称性可知,一般情况下由式(5)得到的矩阵 $\boldsymbol{A}$ 都具有稀疏性.因此,Event_BTM+GPU 模型直接作用于采样过程,用先验知识指导主题采样的同时并不会增加模型的复杂度和推理的难度.在此模型中通过引入事件相似度知识,使得采样过程不会过度偏好个别高频“词条”,充分利用了“长尾词条”的主题贡献,也一定程度上弥补了事件同现稀疏对主题生成带来的负面影响.

3.4 融入事件先验知识的 Event_nBTM 模型

利用 GPU 模型融入事件知识的方法简单直接,并不会改变概率图模型的结构和主题生成过程.然而,Event_BTM 模型由于沿用了 BTM 模型的假设,将共现对与其主题的一致性划上等号,即假设 biterm 中的两个“词条”应具有相同的主题.对于本文中的事件 biterm 来讲,这种假设并不总是合理的.传统 BTM 在短文本中抽取 biterm,由于共现对词汇的距离短,往往处于同一个语义片段内,因此,词对间的主题一致性假设在多数情况下都是成立的.然而,本文中的事件共现对并不一定来源于同一语义片段,共现对内两个事件之间的距离不足以保证其主题的一致性.由此,我们希望利用事件先验知识来调节事件 biterm 中两个事件的主题分配方式.

3.4.1 Event_nBTM 模型

以往的研究表明,先验知识的引入可借助无向的马尔可夫随机场来实现[23-25].我们可在两个事件对应的主题变量间建立关系,用势函数对主题变量的取值进行评估.直观来看,两个事件的相似度越高,主题一致的概率越高.相似事件的主题变量在主题一致时拥有较高的势函数值.由于基于势函数及配分函数等建立的概率分布不具备共轭等便于计算的数学特性,使主题模型的推断过程变得更复杂.因此,本文采用一种更切实可行的方案,分别为每个事件 biterm 设置一个指示变量,其值对应两种可能的状态:

(1) biterm 整体分配同一主题;

(2) biterm 中两个事件的主题独立产生(产生的主题也可能相同).

由此,我们可令指示变量服从伯努利分布.

根据事件先验知识,我们为每个事件 biterm 所对应指示变量的伯努利分布生成超参数,并将此参数加入到模型中以得到相应的指示变量.这种做法可以消除指示变量生成时对事件变量的条件依赖,从而避免网络结构中出现环形结构而无法建模.图 2 给出模型的贝叶斯网络结构.与 Event_BTM 模型比较,此概率图中增加了两类结点.其中,l_b 表示为事件 biterm 设置的隐含指示变量;γ_b 则是根据事件知识得到的超参,用以确定每个指示变量 l_b 所服从的伯努利分布的参数.因此,新的主题模型 Event_nBTM 的生成过程可如下描述:

(1) 对于每个主题 $k\in\{1,2,\cdots,K\}$,生成主题“词条”的多项分布参数 $\varphi_k\sim Dir(\beta)$;

(2) 对于每篇文档 $m\in\{1,2,\cdots,M\}$:

1) 生成文档(事件 biterm 集合)主题多项分布参数 $\theta_m\sim Dir(\alpha)$;

2) 对于文档 m 中的每个共现对 $b(e_i,e_j)$:

① 采样生成指示变量状态 $l_b\sim Bernoulli(\gamma_b)$;

② 若 $l_b=1$,则将 e_i 和 e_j 整体分配同一主题:采样生成主题 $z_b\sim Mult(\theta_m)$,令 $z_{e_i}=z_b$,$z_{e_j}=z_b$;若 $l_b=0$,则 e_i 和 e_j 独立产生主题:采样生成主题 $z_{e_i}\sim$

$Mult(\theta_m)$，采样生成主题 $z_{e_j}\sim Mult(\theta_m)$；

③ 采样生成“词条”$e_i\sim Mult(\varphi_{z_{e_i}})$，$e_j\sim Mult(\varphi_{z_{e_j}})$.

对应其生成过程，我们可得到如下形式的联合概率分布：

$$p(B,Z,L,\theta,\varphi\mid\alpha,\beta,\gamma)=\prod_{k=1}^{K}p(\varphi_k\mid\beta)\times\prod_{m=1}^{M}p(\theta_k\mid\alpha)\times\prod_{b(e_i,e_j)\in B}p(l_b\mid\gamma_b)p(z_{e_i},z_{e_j}\mid\theta_m,l_b)p(e_i\mid\varphi_{z_{e_i}})p(e_j\mid\varphi_{z_{e_j}}) \tag{6}$$

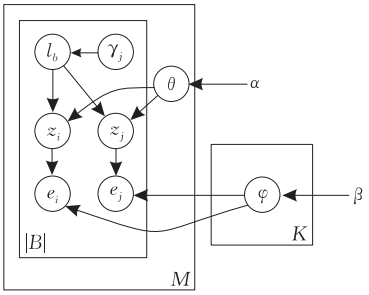

图 2　Event_nBTM 模型

主题生成过程中的重要部分是生成事件对 $b(e_i,e_j)$ 中各个事件的主题，其计算方式如下：

$$p(z_{e_i},z_{e_j}\mid\theta_m,l_b)=\begin{cases}\theta_{m,z_{e_i}}, & l_b=1,z_{e_i}=z_{e_j}\\ 0, & l_b=1,z_{e_i}\neq z_{e_j}\\ \theta_{m,z_{e_i}}\theta_{m,z_{e_j}}, & l_b=0\end{cases} \tag{7}$$

借助指示函数 $I(\cdot)$，可统一表示为如下形式：

$$p(z_{e_i},z_{e_j}\mid\theta_m,l_b)=\theta_{m,z_{e_i}}(\theta_{m,z_{e_j}})^{1-l_b}(I(z_{e_i}=z_{e_j}))^{l_b} \tag{8}$$

式(8)中指示变量的状态决定了当前共现事件是对整个分配主题还是对事件对中两个独立事件分配主题. 设定指示变量 l_b 服从参数为 γ_b 的伯努利分布，其中 γ_b 是由基于事件相似度的先验知识 $sim(e_i,e_j)$ 按式(9)产生的超参数：

$$\gamma_b=\min(\max(sim(e_i,e_j),0)+\sigma,1) \tag{9}$$

其中参数 σ 为非负值.

通过对参数 σ 的调整，可以选择对“共现与主题一致”假设或事件先验知识的偏好. 若 σ 设置为大于 1 的值，则由伯努利分布生成的所有指示变量的值将全部为 1，所有事件对均会被作为整体来生成主题. 此时，当前模型则退化为 Event_BTM 模型. 因此，本文提出的这种 Event_nBTM 模型也可以看作是 Event_BTM 模型的泛化形式. 由此可以看出，尽管本文关注的是基于事件的主题模型，但是我们引入先验知识的方法具有一般性. 因此，这种方法可以应用到其它基于 BTM 的主题模型中.

3.4.2　Event_nBTM 模型推理

与 Event_BTM 模型类似，Event_nBTM 模型仍然使用 Collapsed Gibbs Sampling. 不同的是，我们需要对隐含变量的指示变量进行后验推理. 首先通过吉布斯采样得到隐含主题变量 Z 和指示变量 L 后验概率下的样本，然后利用这些样本结合可观测“事件对”对主题事件分布参数和文档主题分布参数进行后验估计.

(1) 指示变量采样概率

模型中事件对 $b(e_i,e_j)$ 对应的指示变量 l_b 的采样概率为 $p(l_b\mid Z,B,L_{-b};\alpha,\beta,\gamma)$，其计算方法可定义如下：

$$\begin{aligned}p(l_b\mid Z,B,L_{-b};\alpha,\beta,\gamma)&=p(l_b\mid Z,L_{-b};\alpha,\gamma)\\&\propto p(z_{e_i},z_{e_j}\mid l_b,Z_{-b},L_{-b};\alpha)\times p(l_b\mid Z_{-b},L_{-b};\alpha,\gamma)\\&=g(z_{e_i},z_{e_j},l_b)\times p(l_b\mid\gamma_b)\\&=g(z_{e_i},z_{e_j},l_b)\times\gamma_b^{l_b}\times(1-\gamma_b)^{1-l_b}\end{aligned} \tag{10}$$

式中，$g(z_{e_i},z_{e_j},l_b)$ 表示 $p(z_{e_i},z_{e_j}\mid l_b,Z_{-b},L_{-b};\alpha)$.

需要注意的是，由于存在事件对 $b(e_i,e_j)$ 整体和单个事件的两种不同的主题分配方式，因此，在计算后验分布 $p(\theta\mid Z_{-b},L;a)$ 时应分别考虑两种主题生成方式各自使用的频次. 由此，$g(z_{e_i},z_{e_j},l_b)$ 可按如式(11)计算：

$$g(z_{e_i},z_{e_j},l_b)=\int p(z_{e_i},z_{e_j}\mid\theta,Z_{-b},L)p(\theta\mid Z_{-b},L;a)\mathrm{d}\theta$$

$$=\begin{cases}\dfrac{\Gamma(\sum_{k=1}^{K}n_{m,k}^{-b}+\alpha-c_{m,k}^{-b})}{\prod_{k=1}^{K}\Gamma(n_{m,k}^{-b}+\alpha-c_{m,k}^{-b})}\times\dfrac{\prod_{k=1}^{K}\Gamma(n_{m,k}^{-b}+\alpha-c_{m,k}^{-b}+I(k=z_{e_i}))}{\Gamma(\sum_{k=1}^{K}n_{m,k}^{-b}+\alpha-c_{m,k}^{-b}+I(k=z_{e_i}))}, & l_b=1,z_{e_i}=z_{e_j}\\ 0, & l_b=1,z_{e_i}\neq z_{e_j}\\ \dfrac{\Gamma(\sum_{k=1}^{K}n_{m,k}^{-b}+\alpha-c_{m,k}^{-b})}{\prod_{k=1}^{K}\Gamma(n_{m,k}^{-b}+\alpha-c_{m,k}^{-b})}\times\dfrac{\prod_{k=1}^{K}\Gamma(n_{m,k}+\alpha-c_{m,k}^{-b}+I(k=z_{e_i}))}{\Gamma(\sum_{k=1}^{K}n_{m,k}+\alpha-c_{m,k}^{-b}+I(k=z_{e_i}))}, & l_b=0\end{cases} \tag{11}$$

其中：$n_{m,k}$ 表示文档 m 中的事件分配到主题 k 的频次；$c_{m,k}$ 表示文档 m 中以 biterm 方式分配到主题 k 的事件对的频次(若文档 m 中的事件对，其主题变

量 $z_{e_i}=z_{e_j}=k$ 且指示变量 $l_b=1$，则 $c_{m,k}$ 加 1).

借助指示函数，上式可统一表示为如下形式：

$$g(z_{e_i},z_{e_j},l_b)=\frac{n_{m,z_{e_i}}^{-b}-c_{m,z_{e_i}}^{-b}+\alpha}{\sum_{k=1}^{K}(n_{m,k}^{-b}-c_{m,k}^{-b}+\alpha)}\times\left(\frac{n_{m,z_{e_j}}^{-b}-c_{m,z_{e_j}}^{-b}+\alpha+I(z_{e_i}=z_{e_j})}{\sum_{k=1}^{K}(n_{m,k}^{-b}-c_{m,k}^{-b}+\alpha)+1}\right)^{1-l_b}\times(I(z_{e_i}=z_{e_j}))^{l_b} \tag{12}$$

综合式(10)和式(12)，指示变量采样概率的计算可表示如下：

$$p(l_b|Z,B,L_{-b};\alpha,\beta,\gamma)\propto\gamma_b^{l_b}\times(1-\gamma_b)^{1-l_b}\times(I(z_{e_i}=z_{e_j}))^{l_b}\times\frac{n_{m,z_{e_i}}^{-b}-c_{m,z_{e_i}}^{-b}+\alpha}{\sum_{k=1}^{K}(n_{m,k}^{-b}-c_{m,k}^{-b}+\alpha)}\times\left(\frac{n_{m,z_{e_j}}^{-b}-c_{m,z_{e_j}}^{-b}+\alpha+I(z_{e_i}=z_{e_j})}{\sum_{k=1}^{K}(n_{m,k}^{-b}-c_{m,k}^{-b}+\alpha)+1}\right)^{1-l_b} \tag{13}$$

(2) 主题变量采样概率

本文采用联合采样的方案，对事件对所包含两个事件的主题变量 z_{e_i} 和 z_{e_j} 进行联合采样，采样概率为 $p(z_{e_i},z_{e_j}|Z_{-b},B,L;\alpha,\beta)$，其计算方式可表示如下：

$$\begin{aligned}&p(z_{e_i},z_{e_j}|Z_{-b},B,L;\alpha,\beta)\propto\\&p(z_{e_i},z_{e_j}|Z_{-b},B_{-b},L;\alpha)\times p(b|z_{e_i},z_{e_j},Z_{-b},B_{-b},L;\beta)=\\&p(b|z_{e_i},z_{e_j},Z_{-b},B_{-b};\beta)\times p(z_{e_i},z_{e_j}|Z_{-b},L;\alpha)=\\&f(z_{e_i},z_{e_j},e_i,e_j)\times g(z_{e_i},z_{e_j},l_b)\end{aligned} \tag{14}$$

其中后项 $g(z_{e_i},z_{e_j},l_b)$ 的计算已在前面讨论，前项 $f(z_{e_i},z_{e_j},e_i,e_j)$ 可按下述公式进一步计算：

$$\int p(b|z_{e_i},z_{e_j},\varphi)p(\varphi|Z_{-b},B_{-b};\beta)\mathrm{d}\varphi=$$

$$\begin{cases}\dfrac{\Gamma(\sum_{v=1}^{V}n_{z_{e_i},v}^{-b}+\beta)}{\prod_{v=1}^{V}\Gamma(n_{z_{e_i},v}^{-b}+\beta)}\times\dfrac{\prod_{v=1}^{V}\Gamma(n_{z_{e_i},v}+\beta)}{\Gamma(\sum_{v=1}^{V}n_{z_{e_i},v}+\beta)}, & z_{e_i}\neq z_{e_j}\\ \dfrac{\Gamma(\sum_{v=1}^{V}n_{z_{e_i},v}^{-b}+\beta)}{\prod_{v=1}^{V}\Gamma(n_{z_{e_i},v}^{-b}+\beta)}\times\dfrac{\prod_{v=1}^{V}\Gamma(n_{z_{e_i},v}^{-e_j}+\beta)}{\Gamma(\sum_{v=1}^{V}n_{z_{e_i},v}^{-e_j}+\beta)}\times & \\ \dfrac{\Gamma(\sum_{v=1}^{V}n_{z_{e_j},v}^{-b}+\beta)}{\prod_{v=1}^{V}\Gamma(n_{z_{e_j},v}^{-b}+\beta)}\times\dfrac{\prod_{v=1}^{V}\Gamma(n_{z_{e_j},v}^{-e_i}+\beta)}{\Gamma(\sum_{v=1}^{V}n_{z_{e_j},v}^{-e_i}+\beta)}, & z_{e_i}\neq z_{e_j}\end{cases} \tag{15}$$

其中 $n_{z_{e_i},e_i}$ 和 $n_{z_{e_j},e_j}$ 分别表示事件 e_i 和 e_j 分配到主题 z_{e_i} 和 z_{e_j} 的频次. 借助指示函数，式(15)可重新表示为如下形式：

$$f(z_{e_i},z_{e_j},e_i,e_j)=\frac{n_{z_{e_i},e_i}^{-b}+\beta}{\sum_{v=1}^{V}(n_{z_{e_i},v}^{-b}+\beta)}\times\frac{n_{z_{e_j},e_j}^{-b}+\beta+I(z_{e_i}=z_{e_j})}{\sum_{v=1}^{V}(n_{z_{e_j},e_j}^{-b}+\beta)+I(z_{e_i}=z_{e_j})} \tag{16}$$

综合式(12)、式(14)和式(16)，主题变量采样概率的计算可表示如下：

$$\begin{aligned}&p(z_{e_i},z_{e_j}|Z_{-b},B,L;\alpha,\beta)\propto\\&f(z_{e_i},z_{e_j},e_i,e_j)\times g(z_{e_i},z_{e_j},l_b)=\\&\frac{n_{z_{e_i},e_i}^{-b}+\beta}{\sum_{v=1}^{V}(n_{z_{e_i},v}^{-b}+\beta)}\times\frac{n_{z_{e_j},e_j}^{-b}+\beta+I(z_{e_i}=z_{e_j})}{\sum_{v=1}^{V}(n_{z_{e_j},e_j}^{-b}+\beta)+I(z_{e_i}=z_{e_j})}\times(I(z_{e_i}=z_{e_j}))^{l_b}\times\frac{n_{m,z_{e_i}}^{-b}-c_{m,z_{e_i}}^{-b}+\alpha}{\sum_{k=1}^{K}(n_{m,k}^{-b}-c_{m,k}^{-b}+\alpha)}\times\left(\frac{n_{m,z_{e_j}}^{-b}-c_{m,z_{e_j}}^{-b}+\alpha+I(z_{e_i}=z_{e_j})}{\sum_{k=1}^{K}(n_{m,k}^{-b}-c_{m,k}^{-b}+\alpha)+1}\right)^{1-l_b}\end{aligned} \tag{17}$$

(3) 主题分布参数估计

通过吉布斯采样得到主题变量 Z 和指示变量 L 后验概率下的样本后，可结合可观测的 biterm 变量 B，我们可对文档主题分布 θ 和事件主题分布 φ 进行后验估计：

$$\theta_{m,t}=\frac{n_{m,t}-c_{m,t}+\alpha}{\sum_{k=1}^{K}(n_{m,k}-c_{m,k}+\alpha)} \tag{18}$$

$$\varphi_{k,e}=\frac{n_{k,e}+\beta}{\sum_{v=1}^{V}(n_{k,v}+\beta)} \tag{19}$$

其中，$n_m=(n_{m,1},\cdots,n_{m,K})$ 和 $n_k=(n_{k,1},\cdots,n_{k,V})$ 分别表示文档 m 分配到各主题的频次和主题 k 分配给各事件的频次，$c_m=(c_{m,1},\cdots,c_{m,K})$ 表示文档 m 中以事件 biterm 的方式分配到各主题的频次.

4 实 验

4.1 实验数据

本文在中文语料上验证模型. 同文献[36]和文献[37]的设置相似，我们从新浪网上抓取了 2014 年的 10 个专题文档(共计 606 篇)，包括，“云南景谷发生 6.6 级地震”专题(64 篇)，“今年第 9 号台风威马

逊来袭”专题(92 篇),“台湾客机迫降重摔起火”专题(102 篇),“广东遭遇 20 年来最严重登革热疫情”专题(38 篇),“今年第 15 号台风海鸥来袭”(36 篇),“台湾高雄发生气爆事故”(61 篇),“多地再现 H7N9 禽流感病例”(94 篇),“广州公交车爆炸”(30 篇),“杭州发生公交车纵火案”(54 篇),“沪昆高速湖南段发生爆燃事故”(35 篇). 我们未使用公开的文本分类语料或其它主题模型评估语料,主要有以下原因:

(1) 中文主题模型的评估尚未有公开可用的标准语料;

(2) 公开的文本分类语料其主题的语义粒度较粗,用细粒度语义的事件来描述其主题并不适合;

(3) 新浪网中同一专题的文档本身具有较强的相关性,用事件来描述专题比词汇更有意义.

本文使用 Stanford CoreNLP① 对文档进行了分词和依存分析等预处理,事件抽取完成后每篇文档平均 48 个事件. 表 2 列出语料中词和事件的统计结果. 从事件字典的大小及事件词条的总数来看,每个事件平均出现的频率不到 2 次. 由此可见,该语料中出现的事件多为“长尾事件”,基于事件的主题模型所面临的稀疏性问题非常严重.

表 2　语料统计结果

	字典	总数	平均个数/文档
词	17 063	161 264	266
事件	20 866	29 276	48

4.2　实验结果

为验证本文提出方法的有效性,我们对比了 4 种基于事件的主题模型,分别为 Event_LDA(在基本 LDA 基础上使用事件结构作为基本元素)、Event_BTM(在 BTM 基础上使用事件结构作为基本元素)、Event_BTM+GPU(在 Event_BTM 基础上使用 GPU 模型融入事件知识)、Event_nBTM(将事件知识融入主题模型的网络结构).

此外,为验证事件对主题表示的有效性,我们也将基于词的 LDA 模型(Word_LDA)作为基线系统与本文提出的模型进行比较.

实验中,所有主题模型的狄利克雷先验超参数均使用相同的设置,分别设置如下:$\alpha=50/K$,$\beta=0.1$. 此外,相似度阈值参数 $\sigma=0.6$.

4.2.1　内部评估

主题凝聚度(Topic Coherence)和 KL 散度(KL-Divergence)是目前许多主题模型选择的评估指标[26,28],能够从主题中“词条”分布的内聚程度和差异性两个方面反映主题的质量. 本文也选用这两个指标对模型进行自动评估.

(1) 主题凝聚度性能评估

主题凝聚度评估依赖于文档内的“词条”的同现统计,而不是任何外部资源或人工标注. 主题凝聚度得分越高,主题的质量越高. 其计算方法为

$$coherence(V)=\sum_{(v_i,v_j)\in V} score(v_i,v_j,\varepsilon) \quad (20)$$

其中:V 是描述主题的“词条”集合;ε 为平滑因子(通常取 1). 主题凝聚度得分是主题“词条”的点对分布相似度之和.

图 3 给出了各模型在不同主题“词条”数下平均主题凝聚度得分变化的结果. 可以看出,简单使用事件作为基本元素的 LDA 得分最低,而使用 biterm 结构的 3 种模型均有显著提升. 这说明使用 biterm 的二元结构来建模对解决事件稀疏性有着明显的作用. 对比 3 种基于 biterm 的模型,可以发现引入事件知识的两种模型都优于 Event_BTM 模型,当主题词数大于 30 时提升效果比较明显. 与模型 Event_BTM+GPU 相比,Event_nBTM 模型通过引入隐变量,增加了同一 biterm 中两个事件不同主题的可能性,一定程度上提升了主题的凝聚度. 因此,本文采用的事件知识对主题模型的生成有着积极的意义,所提出的两种融入事件知识的模型从共现和语义两个方面较好地解决了事件稀疏问题.

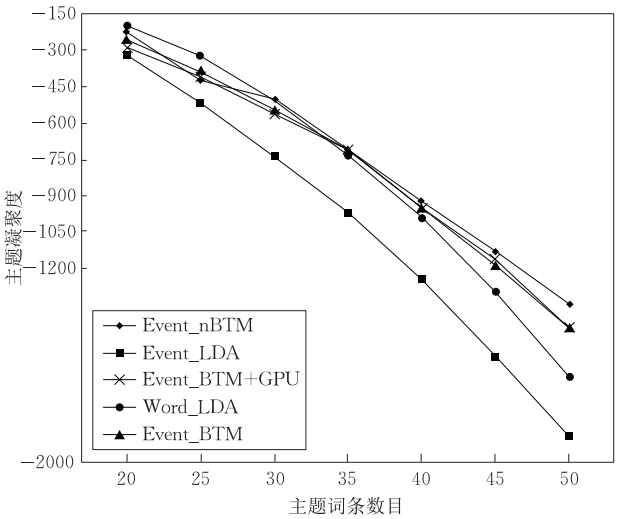

图 3　各个模型不同主题“词条”数下主题凝聚度对比

此外,与 Word_LDA 相比较,当主题词数高于 40 时,3 种基于 BTM 的模型能取得更高的主题凝

① http://stanfordnlp.github.io/CoreNLP/

聚度得分,而在主题词数低于 30 时,其效果略逊于 Word_LDA 模型.造成这一情况的原因可作如下解释:事件在表现形式上的多样化会大大降低共现度,当在主题"词条"数据较少时表现尤为明显;相应地,随着主题"词条"数量的增加,更多在表现形式上不一致但语义相关事件都会出现.这些事件对主题区分有着更大贡献,显著地提升模型的主题凝聚度得分.

(2)KL 散度性能评估

KL 散度用以度量主题区分度,即不同类别间的"词条"分配差异.其值是非对称的.一般取主题对的 KL 散度均值来表示.平均 KL 散度越大,越能区分主题,主题质量越高.单向 KL 散度的计算方法如下:

$$\mathrm{KL}(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{21}$$

表 3 给出了各模型不同主题中"词条"分布的平均 KL 散度.从实验结果来看,效果与主题凝聚度评估的结论是一致的.受事件稀疏性的影响,Event_LDA 得分最低,而基于事件 biterm 的 3 种模型均比简单的 LDA 模型得分有了显著提高.与 Event_BTM 模型相比,融入事件知识后的两种模型均有一些提升.其中 Event_nBTM 的 KL 散度得分最为理想,生成的主题具有更好的区分度.

与 Word_LDA 比较,3 种基于事件 biterm 的模型都取得非常显著的 KL 散度得分.这说明语义信息更丰富的事件对主题的区分有着更明显的作用.

表 3 各模型 KL 散度结果比较

模型	KL 散度
Word_LDA	3.767
Event_LDA	1.614
Event_BTM	7.050
Event_BTM+GPU	7.081
Event_nBTM	7.105

(3)不同主题数目设置的性能评估

为评估模型在不同主题数目时的性能稳定性,我们就不同主题数目的语料进行了比较实验.主题数目取 2～10.图 4 给出了在主题"词条"数固定为 40 时各模型在不同主题个数时的主题凝聚度得分,图 5 则对应各模型的 KL 散度值.

从图中可以看出,在不同主题个数设置下,BTM 模型在处理数据稀疏性问题上比 LDA 模型有着更显著的作用.在引入事件知识后,模型的两个评估指标均有比较明显的提升.因此,本文提出的两种融入事件知识的主题模型有着稳定的性能和显著的优越性.

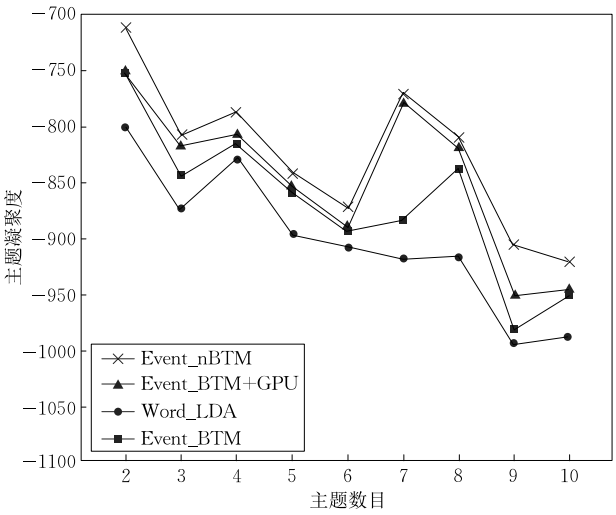

图 4 不同主题数目设置时各模型主题凝聚度对比

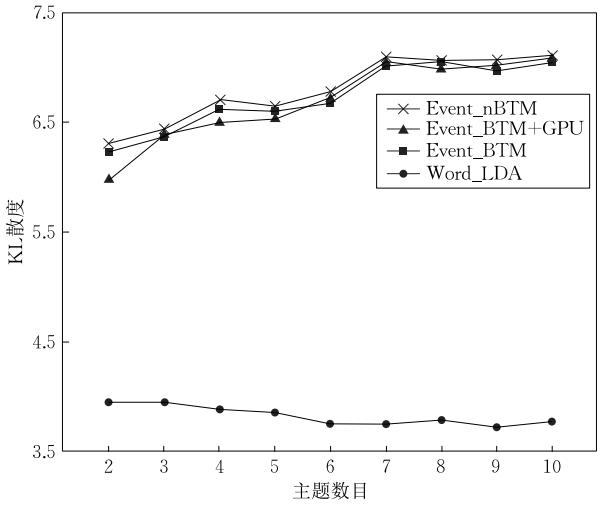

图 5 不同主题数目设置时各模型 KL 散度对比

4.2.2 主题表示样例

基于事件的主题表示比基于词的主题表示具有更好的可解释性.为了更直观地对比两种方式得到的主题表示结果,表 4 和表 5 分别列举了 Event_nBTM

表 4 "台湾客机迫降重摔起火"主题表示

事件	人,死亡 人,罹难 失事,飞机 专家,表示 客机,坠毁 雷雨,交加 航班,失联 飞机,存在,问题 耗尽,燃油 客机,起飞 飞机,失事 风切变,元凶 航班,遭遇,变化 海域,发生,空难 飞机,坠毁 飞机,遭遇,风切变 机型,发生,事故 航班,失事 航空,发生,事故 机组,判断
词	飞机 台湾 航空 复兴 澎湖 机场 失事 人 表示 客机 空难 调查 航班 报道 班机 公司 事故 现场 人员 家属

表 5 "今年第 9 号威马逊台风来袭"主题表示

事件	受,影响 损失,元 威马逊,登陆 人,死亡 做好,工作 确保,安全 气象台,发布,预警 人,受灾 损坏,农房 抗击,台风人,安置 启动,响应 农作物,受灾 威马逊,造成 威马逊,登陆,我国 威马逊,登陆,省区 村民,被困 通讯,中断 转移,人 台风,来袭
词	台风 级 威马逊 登陆 时 沿海 广东 海南 地区 中心 广西 影响 暴雨 18 日 米 预计 19 日 超强 已 附近

模型和 Word_LDA 模型产生的前 20 个主题“词条”(分别对应“台湾客机迫降重摔起火”和“今年第 9 号威马逊台风来袭”两个主题).

不难看出,采用事件表示的主题含义明确,可解释性强.而采用词表示的主题因为语义关联分割后带来了更多的歧义,需要将多个相关的词联系后才能明确主题所对应的语义.例如,在“今年第 9 号威马逊台风来袭”的主题词中出现了“海南”、“广西”、“广东”等词频较高的地名.这些地名与主题密切相关但难以表达主题的明确语义.而对应于基于事件的主题表示,“威马逊,登陆”、“损失,元”等事件可以直观地反映出该主题中“台风登陆”、“带来经济损失”等密切相关的信息.显然,以事件作为主题表示的基本单元对于篇章语义的理解是非常有意义的.

4.2.3　外部评估

主题模型生成每个“词条”在各主题下的概率.因此,主题模型的结果可以应用到文本分类任务中,主题“词条”的选择可以看作特征降维的过程.直观来看,基于事件的主题表示应比词汇具有更强的主题表征能力.本文采用与文献[34]类似的方法来评估各主题模型,分类方法采用 SVM 算法,不同主题模型生成的“词条”作为分类特征,“词条”主题概率作为特征值.实验中使用 libsvm 工具包,核函数为高斯核,实验结果为十折交叉验证后得到的 Accuracy 值.

图 6 给出了各模型在不同主题“词条”数目设置时的分类结果.当主题“词条”数目较少时,基于词特征的分类结果并不理想,随着“词条”数的增加,更多主题特征词的加入对分类效果的提升起到了积极的作用.相比较而言,基于事件特征的分类效果更为稳定,表明本文提出的模型较好地习得具有明显主题特征的事件.

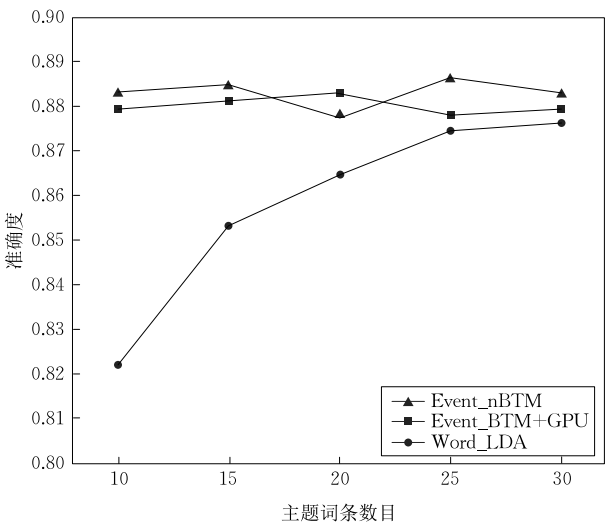

图 6　各个模型不同主题“词条”数下分类结果对比

4.3　结果分析

各模型的对比实验结果表明,基于事件的主题表示是有意义的.从主题表示的结果来看,本文提出的模型也较好地解决了事件作为基本单元所带来的稀疏性问题.分析实验结果,发现仍有几个问题需要在未来的工作中进一步解决和改善:

(1)由于中文语料中大量出现的指代问题(尤其是零指代)使得事件论元缺失的情况比较严重;此外,依存分析的结果使得“同指事件”的表现形式出现多样化,如“人,死亡”和“人,罹难”.本文事件知识的引入一定程度上克服了以上问题所带来的负面影响.采用合理的事件归一化应能对主题生成的质量起到更积极的作用.

(2)部分事件因为中文分词的错误而不可解释,加剧了事件的稀疏度.如“风切变”这一专业词汇,多个文档中都被错误的切分为“风”、“切变”两个词.因而,与“风切变”相关的一些事件无法正常抽取,在一定程度上增加了主题生成的难度.

(3)本文中开放事件抽取采用的是基于规则的方法,其事件抽取的结果依赖于依存分析工具的性能.依存分析的错误增加了长尾事件的数量并直接加剧了事件的稀疏度.因此提升依存分析的性能或采用更有效的开放事件抽取方法对基于事件的篇章语义表示显得尤为重要.我们将中文开放事件抽取作为未来工作的一部分.

5　结论和未来工作

为解决基于词的主题表示可解释性差的问题,本文提出以事件为基本单元表示主题.针对结构化事件的稀疏性问题给主题生成带来的挑战,本文提出两种在基于事件 biterm 的主题模型中融入事件语义先验知识的方法.实验结果表明,基于事件的表示方法明显提高了主题的可读性.本文提出的两个模型,从共现和语义相关两个角度有效地降低了事件稀疏性的影响.

本文主要关注融入事件知识的主题表示,对事件抽取及事件知识的生成并未作更深入的研究.未来的工作主要关注以下 3 个方面:(1)事件作为主题模型的输入对主题的生成有着最直接的影响,研

究开放事件的抽取是非常有意义;(2)本文中的事件知识来源于大规模语料习得的分布式表示,研究知识正确性的判别有助于更好地提高主题的可解释性;(3)在事件主题的生成中加入时序关系和篇章关系以学习对主题的深层语义表示.

参 考 文 献

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

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

[3] Heinrich G. A generic approach to topic models//Proceedings of the European Conference on Machine Learning & Knowledge Discovery in Databases. Bled, Slovenia, 2009: 517-532

[4] Fei Ge-Li, Chen Zhi-Yuan, Liu Bing. Review topic discovery with phrases using the Pólya Urn Model//Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 667-676

[5] Bejan C A. Unsupervised discovery of event scenarios from texts//Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference. Florida, USA, 2008: 124-129

[6] Kitajima R, Kobayashi I. A latent topic extracting method based on events in a document and its application//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Student Session. Portland, Oregon, USA, 2011: 30-35

[7] Ding Xiao, Zhang Yue, Liu Ting, Duan Jun-Wen. Using structured events to predict stock price movement: An empirical investigation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1415-1425

[8] Ng Jun-Ping, Chen Yan, Kan Min-Yen, Li Zhou-Jun. Exploiting timelines to enhance multi-document summarization//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 923-933

[9] Sun Rui, Zhang Yue, Zhang Mei-Shan, Ji Dong-Hong. Event-driven headline generation//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 462-472

[10] Yan Xiao-Hui, Guo Jia-Feng, Lan Yan-Yan, Cheng Xue-Qi. A biterm topic model for short texts//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 1445-1456

[11] Mahmoud H. Polya Urn Models, Texts in Statistical Science. Florida, USA: CRC Press, 2009

[12] Xu Ge, Wang Hou-Feng. The development of topic models in natural language processing. Chinese Journal of Computers, 2011, 34(8): 1423-1436(in Chinese)
(徐戈,王厚峰. 自然语言处理中主题模型的发展. 计算机学报, 2011, 34(8): 1423-1436)

[13] Mei Qiao-Zhu, Zhai Cheng-Xiang. A mixture model for contextual text mining//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 649-655

[14] Mei Qiao-Zhu, Ling Xu, Wondra M, et al. Topic sentiment mixture: Modeling facets and opinions in weblogs//Proceedings of the 16th International Conference on World Wide Web. Banff, Canada, 2007: 171-180

[15] Titov I, McDonlad R. Modeling online reviews with multi-grain topic models//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 111-120

[16] Zhao Xin, Jiang Jing, Yan Hong-Fei, Li Xiao-Ming. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. MIT, USA, 2010: 56-65

[17] Wallach H M. Topic modeling: Beyond bag-of-words//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 977-984

[18] Wang Xue-Rui, Mccallum A, Wei Xing. Topical N-Grams: Phrase and topic discovery, with an application to infromation retrieval//Proceedings of the 2007 Seventh IEEE International Conference on Data Mining. Omaha, USA, 2007: 697-702

[19] Hong Liang-Jie, Yin Da-Wei, Guo Jian, Davison B D. Tracking trends: Incorporating term volume into temporal topic models//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Manchester Grand Hyatt. San Diego, CA, 2011: 484-492

[20] Mei Qiao-Zhu, Cai Deng, Zhang Duo, Zhai Cheng-Xiang. Topic modeling with network regularization//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 101-110

[21] Kataria S, Mitra P, Caragea C, Lee Giles C. Context sensitive topic models for author influence in document networks//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 2274-2280

[22] Andrzejewski D, Zhu X, Craven M. Incoporating domain knowledge into topic modeling via Dirichlet forest priors//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 25-32

[23] Andrzejewski D, Zhu X, Craven M, Recht B. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1171-1177

[24] Xie Peng-Tao, Yang Di-Yi, Xing E P. Incorporating word correlation knowledge into topic modeling//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA, 2015: 725-734

[25] Yang Yi, Downey D, Boyd-Graber J. Efficient methods for incorporating knowledge into topic models//Proceddings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Protugal, 2015: 308-317

[26] Mimno D, Wallach H M, Talley E M, Mccallum A. Optimizing semantic coherence in topic models//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 262-272

[27] Chen Zhi-Yuan, Mukherjee A, Liu Bing, et al. Leveraging multi-domain prior knowledge in topic models//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2071-2077

[28] Chen Zhi-Yuan, Mukherjee A, Liu Bing, et al. Discovering coherent topics using general knowledge//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, USA, 2013: 209-218

[29] Chen Zhi-Yuan, Mukherjee A, Liu Bing, et al. Exploiting domain knowledge in aspect extraction//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1655-1667

[30] Chambers N, Jurafsky D. Unsupervised learning of narrative event chains//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Columbus, Ohio, 2008: 789-797

[31] Hu Zhi-Chao, Rahimtoroghi E, Munishkina L, et al. Unsupervised induction of contingent event pairs from film scenes//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 369-379

[32] Li Pei-Feng, Zhu Qiao-Ming, Zhou Guo-Dong. Argument inference from relevant event mentions in Chinese argument extraction//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 1477-1487

[33] Sun Rui, Wang Zhen-Chao, Ren Ya-Feng, Ji Dong-Hong. Query-biased multi-document abstractive summarization via submodular maximization using event guidance//Proceedings of the 17th International Conference on Weg-Age Information Management. Nanchang, China, 2016: 310-322

[34] Zhang Mu-Yu, Qin Bing, Zheng Mao, et al. Encoding distributional semantic into triple-based knowledge ranking for document enrichment//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 524-533

[35] Grefenstette E, Sadrzadeh M. Experimental support for a categorical compositional distributional model of meaning//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 1394-1404

[36] Zhao Xu-Jian, Yang Chun-Ming, Li Bo, et al. A topic evolution mining algorithm of news text based on feature evolving. Chinese Journal of Computers, 2014, 37(4): 819-832(in Chinese)
(赵旭剑，杨春明，李波等. 一种基于特征演变的新闻话题演化挖掘方法. 计算机学报，2014，37(4)：819-832)

[37] Li Xiang-Dong, Ba Zhi-Chao, Huang Li. News topic mining method based on weighted latent Dirichlet allocation model. Journal of Computer Applications, 2014, 34(5): 1354-1359 (in Chinese)
(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. 计算机应用，2014，34(5)：1354-1359)

SUN Rui, born in 1977, Ph. D. candidate. His main research interests include natural language processing and machine learning.

GUO Sheng, born in 1992, M. S. candidate. His research interest is natural language processing.

JI Dong-Hong, born in 1967, professor, Ph. D. supervisor. His research interests include natural language processing and information retrieval.

Background

This work focuses on event-based topic representation for Chinese open text. Topic model has been widely used to discover the latent topic of text. The topic representation presents the main concepts and semantics of a set of related documents. Most of previous methods take words or phrases as the basic units of topic representation. However, these forms have a poor interpretability, due to the lack of deep semantic information.

In this paper, we regard the structured events as the basic units. We think the events have more abundant semantic information than words or phrases. However, event-based topic models face the data sparseness. This paper proposed two topic models based on Biterm Topic Model. The event semantic knowledge is incorporated into the topic models.

Thanks to the support of Project Event-Chain Model of Chinese Discourse Coherence (Grant No. 61373108). This project aims to construct a model for discourse coherence discrimination based on the concept of event chain which transcends from the traditional entity level to event level. It facilitates the representation mechanism of Chinese discourse semantics, enforces the discourse-level semantic processing. As an important part of this project, this work tries to learn a set of events to represent the topics and the corresponding documents. Furthermore, this paper is supported in part by the State Key Program of National Natural Science Foundation of China (Grant No. 61133012), the National Natural Science Foundation of China (Grant No. 61373056), the National Philosophy Social Science Major Bidding Project of China (Grant No. 11&ZD189).

Our research mainly attempts to learn the Chinese discourse semantic based on event structure and have achieved some promising results involved in event extraction, headline generation, multi-document summarization. In these discourse semantic learning tasks, the results demonstrated the structured events are more promising and effective than traditional words and phrases. We also have paid some attention to topic model in some NLP task, such as word sense induction and sentiment analysis. More details can be found on the author's publications.