



结合主题词嵌入和注意力机制的主题模型

覃婷婷, 刘 峥, 陈可佳

(南京邮电大学 计算机学院, 南京 210023)

摘 要: 社交软件的普及使得从海量数字文本中挖掘有效信息成为一个热点问题, 经典主题模型 LDA 和 LSA 均基于单词共现来捕获主题信息, 忽略了单词间的位置信息。为此, 设计主题与单词间的注意力机制并将主题信息和单词信息融入到 LDA 框架中, 构建一种主题模型 JEA-LDA。该模型通过单词与主题间的注意力机制将单词信息和主题信息融合为特征表示, 用于 LDA 模型的主题提取。实验结果表明, 相比 LDA、DMM 等模型, 该模型的主题一致性和分类性能均较高, 能够取得更好的主题提取效果。

关键词: 主题模型; 单词嵌入; 主题嵌入; 注意力机制; LDA 模型

开放科学(资源服务)标志码(OSID):



中文引用格式: 覃婷婷, 刘峥, 陈可佳. 结合主题词嵌入和注意力机制的主题模型[J]. 计算机工程, 2020, 46(11): 104-108.

英文引用格式: QIN Tingting, LIU Zheng, CHEN Kejia. Topic model combining topic word embedding and attention mechanism[J]. Computer Engineering, 2020, 46(11): 104-108.

Topic Model Combining Topic Word Embedding and Attention Mechanism

QIN Tingting, LIU Zheng, CHEN Kejia

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

[Abstract] With the popularity of social software, mining effective information from massive digital documents has been a hotspot. The classic topic models including LDA and LSA capture topic information based on word co-occurrence and ignore the context information of words. To address the problem, this paper designs an attention mechanism between words and topics, integrates the topic information and word information into the LDA framework, and on this basis constructs a JEA-LDA topic model. The model uses the attention mechanism between words and topics to merge the word information and topic information into feature representation for topic extraction of the LDA model. The experimental results show that compared with LDA, DMM and other models, the proposed model has better performance in topic coherence and classification tasks, and improves the topic extraction results.

[Key words] topic model; word embedding; topic embedding; attention mechanism; LDA model

DOI:10.19678/j.issn.1000-3428.0055952

0 概述

随着互联网行业的快速发展, 文档数据急剧增加, 从文本数据中发现潜在的主题信息也变得更加困难。经典主题模型如 LDA^[1] 和 sentenceLDA^[2] 通常利用文档或者句子级别的单词共现来构成主题, 根据简单的词袋模型捕获单词之间的语义信息, 但是, 该方法忽略了有价值的单词序列信息^[3]。目前, 研究人员提出了引入单词嵌入和主题嵌入的主题模型 LTE(Latent Topic Embedding)^[3], 其将单词嵌入

和主题模型集成到一个框架中。单词嵌入模型^[4]将单词映射到分布式表示中, 其主要关注小滑动窗口内的单词共现, 这使得单词嵌入可以捕获单词序列的信息。但是, 现有的单词嵌入模型通常只关注单词上下文的语义信息, 并未充分了解文本的主题。

目前, 学者们关于主题建模和单词嵌入进行了较多研究。LDA 是用离散数据集合(如文本语料库)建立的生成概率模型^[1], LDA 及其变体已广泛应用于内容推荐^[5-6]、趋势检测^[7-8]以及用户概况分析^[9-10]等应用中。Bigram 主题模型^[11-12]为了减轻

基金项目: 南京邮电大学引进人才科研启动基金(NY215045); 南京邮电大学国家自然科学基金孵化项目(NY219084)。

作者简介: 覃婷婷(1994—), 女, 硕士研究生, 主研方向为自然语言处理; 刘 峥(通信作者), 讲师、博士; 陈可佳, 副教授、博士。

收稿日期: 2019-09-08 **修回日期:** 2019-11-05 **E-mail:** zliu@njupt.edu.cn

LDA 主题模型词袋假设的负面影响,为每一对主题的单词创建多项式分布,这导致其计算成本大幅增加。主题联合词向量模型^[13]通过对单词和主题向量进行线性变换得到最终的词向量。文献[14]将主题模型应用于文档检索,在一定程度上提高了文档检索的效果。文献[4]提出了 Skip-gram 模型的几个扩展模型,提高了向量的质量和训练速度。文献[15]将主题建模的结果输入单词嵌入模型以学习主题词嵌入,但是其并非整合主题建模和单词嵌入。文献[16]基于 LDA 主题模型引入深度神经网络模型 LSTM (Long Short-Term Memory),建立了 LLA (Latent LSTM Allocation) 模型。LLA 模型通过 LSTM 预测每个单词主题的生成概率,使得 LDA 模型的超参数减少,同时利用了上下文的文本信息。但是,LLA 模型用 LSTM 对主题和单词进行嵌入,并且忽略了单词与主题之间的相互关系。在本文模型中,将通过引入注意力机制的方法来解决这一问题。

主题模型可以了解文本的主题信息从而捕获文本的主题分布,使得用户可以较容易地获取文本的主要内容,而单词嵌入可以在一个小的滑动窗口内捕获单词的语义信息,并将单词表示成一个较低维度的分布,这使得衡量单词间距离的难度降低。鉴于主题模型和单词嵌入的优点和缺点,本文使用 LDA 模型作为主要框架,通过注意力机制将主题嵌入和单词嵌入融合到 LDA 模型中,在此基础上,构建一种 JEA-LDA (Joint Embedding and Attention for Latent Dirichlet Allocation) 模型。在文本的生成过程中,本文假设文档中观察到的单词的主题可以通过 2 个通道生成,一个是多项式分布,另一个是基于主题嵌入和单词嵌入。此外,在 JEA-LDA 模型中,针对主题和单词建立注意力机制,获取主题与单词间的相互关系。在训练单词嵌入和主题嵌入的过程中,学习注意力分数,以确保在给定文本中与文本主题相关的单词的权重高于不相关单词的权重,从而使得主题嵌入和单词嵌入的信息将影响主题建模的结果,而主题分布又将影响单词嵌入和主题嵌入的训练。

1 JEA-LDA 主题模型

经典主题模型 LDA 利用单词实例的共现来提取文本主题,但是其忽略了单词间的位置关系。在本文中,通过将主题词嵌入融入到 LDA 主题模型中来预测文本中每个单词的主题,同时本文在主题词嵌入模型中引入注意力机制,计算每个单词的重要性分数,利用单词的重要性分数和主题词嵌入来预测下一个单词的主题。

1.1 模型框架

本文 JEA-LDA 主题模型的贝叶斯网络示意图如图 1 所示。

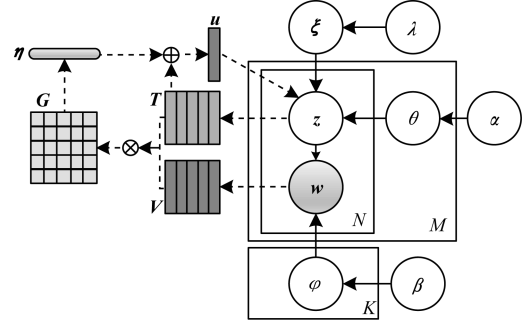


图 1 JEA-LDA 主题模型的贝叶斯网络示意图

Fig. 1 Bayesian network schematic diagram of JEA-LDA topic model

JEA-LDA 主题模型以 LDA 模型为主题框架,融合单词嵌入和主题嵌入并引入注意力机制。由图 1 可以看出,JEA-LDA 主题模型与 LDA 主题模型结构相似,不同之处在于,JEA-LDA 主题模型添加了一个决定参数 λ ,表示主题生成的来源,JEA-LDA 还结合了单词嵌入和主题嵌入结构,并在单词和主题之间添加注意力机制,用来捕获主题与单词之间的相互作用关系。在图 1 中, V 表示短文本单词序列所组成的单词嵌入矩阵,为被预测单词的前导单词序列, T 表示连续单词的主题嵌入矩阵,为被预测单词的前导单词的主题序列。嵌入矩阵的每一列表示一个单词嵌入或主题嵌入,主题嵌入和单词嵌入的长度保持一致。

本文将单词和主题嵌入到同一个维度中,主题嵌入矩阵 $T = \{t_1, t_2, \dots, t_L\}$,单词嵌入矩阵 $V = \{v_1, v_2, \dots, v_L\}$,其中, L 表示窗口大小。假设将单词和主题嵌入到 P 维空间中,则 $T \in \mathbb{R}^{P \times L}$, $V \in \mathbb{R}^{P \times L}$ 。本文通过式(1)计算注意力矩阵 G :

$$G = T^T \cdot V / \hat{G} \quad (1)$$

其中, \hat{G} 表示大小为 $L \times L$ 的归一化矩阵,其每一个元素对应单词嵌入和主题嵌入的 l_2 范数,即 $\hat{g}_{n,n} = \|t_n\|_2 \cdot \|v_n\|_2$ 。

为了捕获连续单词序列(如短语)的相对空间位置信息,本文在注意力的计算过程中引入一个非线性函数 ReLU。特别地,本文考虑一个长度为 $2r+1$ 、中心词为第 n 个单词的单词序列,用注意力矩阵 G 的局部矩阵 $G_{n-r:n+r}$ 来计算主题-短语的注意力分数。本文通过式(2)在第 n 个短语与主题间学习更高级的注意力分数:

$$s_n = \text{ReLU}(G_{n-r:n+r} W_1 + b_1) \quad (2)$$

其中, $W_1 \in \mathbb{R}^{2r+1}$ 和 $b_1 \in \mathbb{R}^1$ 是需要学习的参数。第 n 个短语的最大注意力分数为 $m_n = \max\text{-pooling}(s_n)$, m 是一个长度为 L 的向量。整个单词序列的注意力分数如式(3)所示:

$$\eta = \text{softmax}(m) \quad (3)$$

其中,softmax 函数为 $\eta_n = \frac{\exp(m_n)}{\sum_{n'} \exp(m_{n'})}$ 。本文通过

用注意力分数表示单词的重要程度,对单词序列的主题嵌入进行加权,从而获得待预测单词的主题嵌入表示,如式(4)所示:

$$\mathbf{u} = \sum_{n=1}^L \eta_n \mathbf{t}_n \quad (4)$$

本文用交叉熵来衡量主题表示的概率,即式(4)中 \mathbf{u} 为待预测单词 w 的概率,如式(5)所示:

$$p(z_w | \mathbf{V}, \mathbf{T}) = \text{CE}(\mathbf{Z}_w, f(\mathbf{u})) \quad (5)$$

其中, \mathbf{Z}_w 表示待预测单词主题 z_w 的 one-hot 向量, z_w 表示待预测单词 w 的主题, CE 表示交叉熵函数。 $f(\mathbf{u}) = \text{softmax}(\mathbf{u}')$, $\mathbf{u}' = \mathbf{W}_2 \mathbf{u} + \mathbf{b}_2$ 。 $\mathbf{W}_2 \in \mathbb{R}^{K \times P}$, $\mathbf{b}_2 \in \mathbb{R}^{K \times P}$, K 表示主题个数。在基于注意力的单词嵌入和主题嵌入中,本文模型需要求解的参数有 $\sigma = \{\mathbf{V}, \mathbf{T}, \mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2\}$, 它们将在端到端的学习过程中被训练。对于整个主题词嵌入模型中的全局主题嵌入和单词嵌入的矩阵,本文用预训练的单词嵌入来初始化单词嵌入矩阵,对于不在词汇表中的单词和主题,本文采用随机初始化的方式进行初始化。

在 JEA-LDA 主题模型中,本文首先根据狄利克雷分布先验参数 α 和 β 获取参数文档-主题分布 θ 和主题-词分布 ϕ ; 然后根据多项式分布 $\text{Multi}(\theta)$ 和主题词嵌入模型为每一篇文档的每一个单词选定主题;最后根据多项式分布 $\text{Multi}(\phi)$ 为每一篇文档逐步生成单词。JEA-LDA 模型的生成过程如算法 1 所示。

算法 1 JEA-LDA 模型生成算法

输入 文本数据集 $D = \{d_1, d_2, \dots, d_M\}$

输出 文档-主题分布 θ , 主题-词分布 ϕ

1. for $k = 1$ to K do

2. 根据狄利克雷先验分布抽样主题-词分布 $\phi_k \sim \text{Dir}(\beta)$;

3. end for

4. for each 文档 $d \in D$ do

5. 根据狄利克雷先验分布抽样文档-主题分布 $\theta_d \sim \text{Dir}(\alpha)$;

6. for each 单词 $w \in d$ do

7. 根据伯努利分布抽样一个决定参数 $\xi_w \sim \text{Ber}(\lambda)$;

8. 根据文档-主题分布和主题词嵌入模型的预测概率为单词 w 抽样一个主题 $z_w \sim (1 - \xi_w) \text{Multi}(\theta_d) + \xi_w p(z_w | \mathbf{V}, \mathbf{T})$;

9. 根据主题-词分布抽样一个单词 $w \sim \text{Multi}(\phi_{z_w})$;

10. end for

11. end for

12. return 文档-主题分布 θ , 主题-词分布 ϕ

1.2 模型参数推导

在 JEA-LDA 模型中,单词 w 的概率可描述为 $p(w | \alpha, \beta, \lambda, \sigma)$, 其目标是最大化单词 w 的概率。在理想情况下,可以通过最大化 $p(w | \alpha, \beta, \lambda, \sigma)$ 来计算 σ 的最优值。但是,直接计算 $p(w | \alpha, \beta, \lambda, \sigma)$ 非常困难,因此,本文计算后验概率 $p(w, \xi, z | \alpha, \beta, \lambda, \sigma)$, 如式(6)所示:

$$\begin{aligned} p(w, \xi, z | \alpha, \beta, \lambda, \sigma) &= \\ p(\xi | \lambda) p(z | \alpha, \xi, \sigma) p(w | z, \beta) &= \\ (1 - \lambda)^A \lambda^B \prod_{d=1}^D \prod_{w \in d} p(z_w | \sigma)^{I(\xi=1)} \cdot \\ \left(\frac{\Gamma\left(\sum_{k=1}^K \alpha\right)}{\sum_{k=1}^K \Gamma(\alpha)} \right)^M \prod_{d=1}^D \frac{\sum_{k=1}^K \Gamma(E_{d,k} + \alpha)}{\Gamma\left(\sum_{k=1}^K (E_{d,k} + \alpha)\right)} \cdot \\ \left(\frac{\Gamma\left(\sum_{v=1}^V \beta\right)}{\sum_{v=1}^V \Gamma(\beta)} \right)^M \prod_{k=1}^K \frac{\sum_{v=1}^V \Gamma(F_{k,v} + \beta)}{\Gamma\left(\sum_{v=1}^V (F_{k,v} + \beta)\right)} \end{aligned} \quad (6)$$

其中, $E_{d,k}$ 表示文档 d 中属于主题 k 的单词个数, $F_{k,v}$ 表示文档数据集中属于主题 k 的单词 v 的个数, $\Gamma(\cdot)$ 表示 Gamma 函数, A 表示通过伯努利分布生成的 0 的数量, B 表示通过伯努利分布生成的 1 的数量。根据贝叶斯规则,为文档 d 的单词 w 指定主题 k 的概率如式(7)所示:

$$\begin{aligned} p(z_{d,w} = k, \xi_{d,w} | w, z_{\neg d,w}, \alpha, \beta, \lambda, \sigma) &= \\ p(\xi_{d,w} | \lambda, \xi_{\neg d,w}) p(z_{d,w} = k | w, z_{\neg d,w}, \xi_{d,w}, \alpha, \beta, \sigma) &= \\ (1 - \lambda)^A \lambda^B \prod_{w \in d, \xi_w = 1} p(z_w | \sigma) \cdot \\ \frac{E_{d,k} + \alpha}{\sum_{k'=1}^K (E_{d,k'} + \alpha)} \prod_{w \in d} \frac{F_{d,w} + \beta}{\sum_{w' \in W} (F_{d,w'} + \beta)} \end{aligned} \quad (7)$$

本文根据式(7)整合 ξ_w , 如式(8)所示:

$$\begin{aligned} p(z_{d,w} = k | w, z_{\neg d,w}, \xi_{\neg d,w}, \alpha, \beta, \lambda, \sigma) &= \\ (1 - \lambda)^A \frac{E_{d,k} + \alpha}{\sum_{k'=1}^K (E_{d,k'} + \alpha)} \prod_{w \in d} \frac{F_{d,w} + \beta}{\sum_{w' \in W} (F_{d,w'} + \beta)} + \\ \lambda \prod_{w \in d} p(z_w | \sigma) \end{aligned} \quad (8)$$

本文利用式(8)采样每篇文档中每个单词的主题,重复执行,直至收敛。接下来则考虑单词嵌入和主题嵌入的优化过程。在主题词嵌入的过程中,对于每个短文本 d 的单词 w 的主题,本文用单词 w 前面的单词序列预测 w 的主题。因此,目标函数可以建立如下:

$$\max \sum_{d=1}^D \sum_{w=1}^W p(z_{d,w} | \mathbf{V}, \mathbf{T}) \quad (9)$$

根据上述分析,可以用蒙特卡罗 EM 算法来推导 JEA-LDA 模型的参数,如算法 2 所示。应用该算法可以获得本文模型的参数,如文档-主题分布 θ 和主题-词分布 ϕ 。

算法 2 蒙特卡罗 EM 算法

输入 文本数据集 D

输出 文档-主题分布 θ , 主题-词分布 ϕ

1. 初始化单词嵌入矩阵 \mathbf{V} 和主题嵌入矩阵 \mathbf{T} ;

2. 为每篇文档的每个单词随机指派一个主题;

3. repeat
4. E-Step;
5. for each 文档 $d \in D$ do
6. for each 单词 $w \in d$ do
7. 根据主题词嵌入模型计算 $p(z_{d,w} | \sigma)$;
8. 根据式(8)获取主题 $z_{d,w}$;
9. end for
10. end for
11. M-Step;
12. 用随机梯度下降法优化主题词嵌入模型参数 σ ;
13. until 收敛

在算法 2 中,第 1 行首先对主题词嵌入模型进行初始化,本文用预测训练的单词嵌入初始化单词嵌入矩阵,对于不在词汇表中的单词和主题,本文采用均匀分布进行初始化。第 2 行对每一篇文档随机指派一个主题。第 7 行根据主题词嵌入模型的前向过程预测单词 w 主题为 $z_{d,w}$ 的概率。第 8 行根据式(8)指定单词 w 的主题 $z_{d,w}$ 。第 12 行用随机梯度下降法求解主题词嵌入模型的参数。

假设算法 2 的最大迭代次数为 H ,语料库中文本数量为 M ,每篇文档的平均单词数量为 N ,则 JEA-LDA 模型的时间复杂度为 $O(HMN)$ 。

2 实验结果与分析

2.1 实验数据集

本次实验采用搜狗实验室(<http://www.sogou.com/labs/>)的新闻数据集来预训练单词嵌入,使用爬取自新浪微博的文本数据集来评估文本主题质量,该数据集包括 679 823 条文本数据,每条文本数据包含 100 个~200 个单词。

2.2 对比模型

本次实验的对比模型具体如下:

1) LDA^[1],经典主题模型,直接用 LDA 对文本数据集提取主题。

2) DMM^[17],Dirichlet 多项式混合模型,其主要思想是假设每篇文本仅有一个主题。

3) LF-DMM^[18],DMM 的改进模型,其在 DMM 模型中引入了外部词向量来补充单词间的关系。

4) LF-LDA^[18],LDA 的改进模型,其在 LDA 模型中引入了外部词向量来补充单词间的关系。

2.3 主题一致性评估

本节将比较每个模型的主题质量。采用 PMI (Pointwise Mutual Information)^[19]来衡量主题的一致性,PMI 已被证明是一种有效的主题质量衡量标准^[20]。给定一个主题 k 及其前 T 个单词 $W_k = \{w_1^k, w_2^k, \dots, w_T^k\}$ (即具有最高概率的前 T 个单词), $f(w)$ 表示单词 w 的文档频率, $f(w_i, w'_i)$ 表示单词 w_i, w'_i 同时出现在同一篇文档中的频率。主题 k 的 PMI 分数如下:

$$\text{PMI}(k, W^k) = \sum_{i=1}^T \sum_{j=1}^i \text{lb} \left(\frac{f(w_i^k, w_j^k) + \varepsilon}{f(w_i^k)f(w_j^k)} \right) \quad (10)$$

PMI 得分越高,模型学习的主题一致性越好,即模型性能越高。

图 2 所示为微博文本数据在每个对比模型上的主题一致性 PMI 分数,其中,使用每个主题的前 20 个单词分别计算 PMI 分数。从图 2 可以看出,本文 JEA-LDA 模型相较于其他模型能够取得更好的一致性效果。

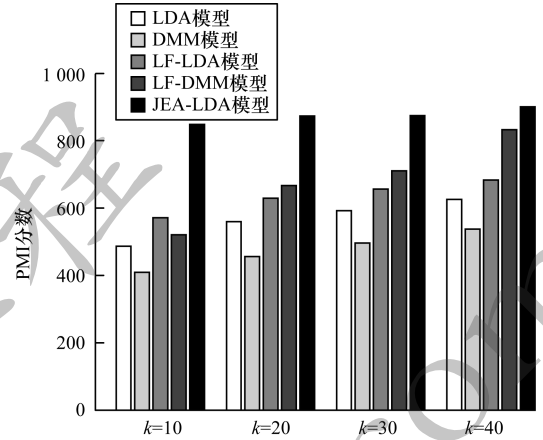


图 2 5 种模型的 PMI 分数对比结果

Fig. 2 Comparison results of PMI score of five models

2.4 分类实验

在本文的对比主题模型中,可以获得模型的文档-主题分布 θ ,因此,可以用通用分类器对文本进行分类,以测试文本主题分布的效果,本次实验采用 SVM 分类器。主题之间的区分度越高,文本主题的分布越合理,分类效果越好,模型的学习能力越高。

本文采用精度(P)、召回率(R)和 F1 值作为每种模型的分类评价指标,5 种模型分类结果对比如表 1~表 3 所示,其中最优结果加粗表示。

表 1 5 种模型分类精度对比

Table 1 Comparison of classification precision of five models

模型	P			
	$k=10$	$k=20$	$k=30$	$k=40$
LDA	0.516	0.567	0.507	0.584
DMM	0.782	0.854	0.811	0.793
LF-LDA	0.669	0.809	0.701	0.849
LF-DMM	0.845	0.733	0.823	0.801
JEA-LDA	0.854	0.859	0.886	0.890

表 2 5 种模型分类召回率对比

Table 2 Comparison of classification recall of five models

模型	P			
	$k=10$	$k=20$	$k=30$	$k=40$
LDA	0.580	0.615	0.545	0.600
DMM	0.775	0.840	0.805	0.805
LF-LDA	0.650	0.815	0.765	0.825
LF-DMM	0.835	0.800	0.820	0.790
JEA-LDA	0.826	0.827	0.855	0.850

表 3 5 种模型的分类 F1 值对比

Table 3 Comparison of classification F1 value of five models

模型	F1 值			
	k = 10	k = 20	k = 30	k = 40
LDA	0.543	0.587	0.522	0.574
DMM	0.777	0.839	0.801	0.795
LF-LDA	0.650	0.797	0.731	0.811
LF-DMM	0.839	0.764	0.796	0.757
JEA-LDA	0.840	0.843	0.870	0.869

从表 1~表 3 可以看出,本文模型通过引入单词嵌入和主题嵌入,在一定程度上改善了主题模型的分类性能。

3 结束语

本文将主题嵌入和单词嵌入融合到 LDA 主题模型中,在主题和单词之间建立注意力机制,获取主题与单词间的相互关系。在训练单词嵌入和主题嵌入的过程中学习注意力分数,以确保在给定文本中与文本主题相关的单词的权重高于不相关单词的权重。实验结果表明,主题嵌入和单词嵌入相结合能够改善主题提取的效果。下一步将在本文研究的基础上,考虑短文本数据稀疏问题,针对短文本的主题提取和注意力机制进行分析和研究。

参考文献

- [1] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2012, 3(1): 993-1022.
- [2] JO Y, OH A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2011: 259-268.
- [3] JIANG Oi, SHI Lei, LIAN Rongzhong, et al. Latent topic embedding[C]//Proceedings of COLING'16. Osaka, Japan: ACL, 2016: 189-206.
- [4] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[EB/OL]. [2019-08-26]. <https://arxiv.org/pdf/1310.4546.pdf>.
- [5] KRESTEL R, FANKHAUSER P, NEJDL W. Latent Dirichlet allocation for tag recommendation[C]//Proceedings of the 3rd ACM Conference on Recommender Systems. New York, USA: ACM Press, 2009: 198-218.
- [6] WAN Pingyu. Design and implementation of recommendation system based on heterogeneous social network relationship and topic model[D]. Beijing: Beijing University of Posts and Telecommunications, 2019. (in Chinese)
万坪禹. 基于异构社交网络关系和主题模型的推荐系统的设计与实现[D]. 北京: 北京邮电大学, 2019.
- [7] LAU J, COLLIER N, BALDWIN T. On-line trend analysis with topic models; #twitter trends detection topic model online[EB/OL]. [2019-08-26]. <http://pdfs.semanticscholar.org/f169/15e4e8d0361b8e577b2123ba4a36a25032ba.pdf>.
- [8] YU Xuewei. Analysis and application of network public opinion based on topic model[D]. Xiamen: Xiamen University, 2017. (in Chinese)
于学伟. 基于主题模型的网络舆情分析及其应用研究[D]. 厦门: 厦门大学, 2017.
- [9] MCCALLUM A, WANG X, CORRADA-EMMANUEL A. Topic and role discovery in social networks with experiments on enron and academic email[J]. Journal of Artificial Intelligence Research, 2007, 30: 249-272.
- [10] GAO Zefeng, WANG Bang, XU Minghua. Event recommendation based on topic model analysis and user long-and short-term interest[J]. Journal of Chinese Computer Systems, 2018, 39(4): 625-630. (in Chinese)
高泽锋, 王邦, 徐明华. 基于主题模型分析与用户长短兴趣的活动推荐[J]. 小型微型计算机系统, 2018, 39(4): 625-630.
- [11] WALLACH H M. Topic modeling: beyond bag-of-words[EB/OL]. [2019-08-26]. <http://people.ee.duke.edu/~lcarin/icml2006.pdf>.
- [12] LI Siyu. Research on semantic mining of short texts based on topic model and word vector[D]. Taiyuan: Taiyuan University of Technology, 2018. (in Chinese)
李思宇. 基于主题模型和词向量的短文本语义挖掘研究[D]. 太原: 太原理工大学, 2018.
- [13] WU Xukang, YANG Xuguang, CHEN Yuanyuan, et al. Topic combined word vector model[J]. Computer Engineering, 2018, 44(2): 233-237, 270. (in Chinese)
吴旭康, 杨旭光, 陈园园, 等. 主题联合词向量模型[J]. 计算机工程, 2018, 44(2): 233-237, 270.
- [14] REN Pengcheng. Research and implementation of intelligent full text retrieval system based on topic ranking and recommendation[D]. Zhengzhou: Zhengzhou University, 2018. (in Chinese)
任鹏程. 基于主题排序与推荐的智能全文检索系统研究与实现[D]. 郑州: 郑州大学, 2018.
- [15] YANG L, LIU Z, CHUA T S, et al. Topical word embeddings[EB/OL]. [2019-08-26]. http://nlp.csai.tsinghua.edu.cn/~lzy/publications/aaai2015_twe.pdf.
- [16] ZAHEER M, AHMED A, SMOLA A J. Latent LSTM allocation joint clustering and non-linear dynamic modeling of sequential data[EB/OL]. [2019-08-26]. <http://proceedings.mlr.press/v70/zaheer17a/zaheer17a.pdf>.
- [17] YIN Jianhua, WANG Jianyong. A Dirichlet multinomial mixture model-based approach for short text clustering[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2014: 1223-1256.
- [18] NGUYEN D Q, BILLINGSLEY R, DU L, et al. Improving topic models with latent feature word representations[J]. Transactions of the Association for Computational Linguistics, 2015, 3: 299-313.
- [19] LAU J H, BALDWIN T. The sensitivity of topic coherence evaluation to topic cardinality[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics. Osaka, Japan: ACL, 2016: 148-156.
- [20] FANG A J, MACDONALD C, OUNIS I, et al. Topics in tweets: a user study of topic coherence metrics for twitter data[M]//XIAO G, SHAN W, SYSTEMS O, et al. Lecture notes in computer science. Berlin, Germany: Springer, 2016: 492-504.