



基于对话内容的交互型文本会话主题挖掘

彭杰¹, 石永革¹, 高胜保²

(1. 南昌大学信息工程学院, 江西 南昌 330029;

2. 中国电信股份有限公司江西分公司, 江西 南昌 330029)

摘要: 传统的主题挖掘模型一般仅从交互型文本中挖掘出文档主题, 为了能够从中挖掘出会话主题并提高挖掘模型的普适性, 提出了一种基于对话内容的交互型文本会话主题生成模型。首先通过分析交互型文本的特征, 基于主题树的概念, 定义了一个 5 层结构的对话生成树。以此为基础, 再基于 LDA 构建会话主题生成模型(ST-LDA)。最后采用吉布斯抽样法对 ST-LDA 进行推导, 得到会话主题及其分布概率。使用实际数据进行验证, 结果表明, ST-LDA 模型可以从交互型文本中有效地挖掘出会话主题。此外, 成果可以降低分类算法的复杂度, 回溯主题—参与者关联关系, 具有较好的普适性。

关键词: 交互型文本; 对话内容; 会话主题挖掘; 对话生成树; LDA

中图分类号: TP181

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016238

Session topic mining for interactive text based on conversational content

PENG Jie¹, SHI Yongge¹, GAO Shengbao²

1. Information Engineering College, Nanchang University, Nanchang 330029, China

2. Jiangxi Branch of China Telecom Co., Ltd., Nanchang 330029, China

Abstract: Traditional theme mining model generally digs out the document theme from the interactive text only. In order to explore the session topic and improve the universality of mining model, a kind of interactive text session topic generation model based on the content of the dialogue was put forward. Firstly, by analyzing the characteristics of interactive text and based on the concept of topic tree, a dialog spanning tree was defined with a five-layer structure. Based on this and LDA, the model of session topic generation (ST-LDA) was built. At last, Gibbs sampling method was adopted to deduce the ST-LDA and obtaining session topic and its distribution probability. The results show that the ST-LDA model can dig out a session topic effectively from the interactive text. Besides, the results can reduce the complexity of the classification algorithm and can be back to the theme—participants association. It also has a good universality.

Key words: interactive text, conversation content, session topic mining, dialog spanning tree, latent Dirichlet allocation

收稿日期: 2016-06-20; 修回日期: 2016-09-10

通信作者: 石永革, ncsyg@163.com

基金项目: 国家自然科学基金资助项目 (No.61163005); 江西省科技计划基金资助项目 (No.2014ZBBE50008)

Foundation Items: The National Natural Science Foundation of China (No.61163005), Science and Technology Program of Jiangxi Province of China (No.2014ZBBE50008)



1 引言

随着移动互联网的迅速发展和智能终端用户数的快速增长,短信、QQ、微信等即时信息交互工具得到了广泛应用。这些应用会产生大量的对话文本,其中隐含有许多值得关注的信息,诸如商业机会、社情舆论、犯罪信息等。对这些信息进行的主题挖掘,可以广泛用于商品精准营销、产品定向开发、舆情分析、罪犯追踪等。对话文本的主要特征包括多角色交互、多主题交织、表达口语化等,其中交互是其根本特征,因此也可称为交互型文本。如果将交互型文本简单地视为多个单条文档的集合,不考虑交互特征,仅基于关键词挖掘出的主题属于文档主题。基于交互性特征再提取出的主题可称为对话主题,以对话主题为基础即可进一步挖掘会话主题。

传统的主题挖掘模型已经比较成熟,典型的如参考文献[1-3]。然而,只能从交互型文本中挖掘出文档主题,难以获知它产生的影响和效果,而这往往是需要特别关注的重要信息。为此,近年来开始研究如何基于对话的交互性特征,从交互型文本中自动挖掘对话主题,并已有若干研究成果。如参考文献[4]提出了一种基于主题森林结构的自组织对话管理模型,其中提出了主题树的概念,但仅适用于人机对话系统,不属于主题挖掘。参考文献[5]可以从对话文本中挖掘出特定关注主题的相关文档,本质上仍属于文档主题挖掘,其主要贡献在于将 LDA (latent Dirichlet allocation) 模型应用于交互型文本。参考文献[6]基于对话的时序关系生成多个主题线索树,可以从交互型文本中析出含有潜在对话主题的主题句,而不是挖掘出对话主题。参考文献[7]基于多项加权(词语出现频率、语句长度和出现位置、对话者等),从对话文本中抽取关键词并进行排序,只适用于纯对话文本的标签推荐。参考文献[8]可以基于微博的特定标识,挖掘微博的对话主题,普适性较差。

综上所述,参考文献[1-8]对于本文具有一定的参考价值,但基本只针对交互型文本实现文本主题或对话主题挖掘,不能实现会话主题挖掘。此外,上述研究成果都不是基于文本内容的关联关系来挖掘主题,必然影响其综合性能(准确性、完整性、普适性)。

为了实现从交互型文本中挖掘出会话主题,以 LDA 模型为基础,提出了一种基于对话内容的交互型文本会话主题生成模型,可以从交互型文本中有效地挖掘出会话主题,并同时挖掘出文档主题、对话主题。为了准确完整地描

述交互型文本的多层主题结构,基于主题树的概念和交互型文本的特征,提出了“文本—会话主题—对话主题—文档主题—词语”的 5 层结构对话生成树。然后基于对话生成树和 LDA 模型,提出交互型文本会话主题生成(session topic-latent Dirichlet allocation, ST-LDA)模型。接下来应用吉布斯抽样(Gibbs sampling)法对 ST-LDA 模型进行推导,实现交互型文本的会话主题挖掘。最后基于实际的对话数据,选用合适的文本分类算法,对本文成果进行了验证。

2 相关工作

2.1 LDA 模型

主题是一段文本(一篇文章、一段话、一个句子等)所表达的中心思想。主题模型是一种使用概率产生式模型来挖掘文本主题的方法。LDA 模型则是目前主题建模的一个标准,非常适于识别大规模文档集或语料库中潜藏的主题信息,其建立的是一种文档—主题—词的 3 层贝叶斯模型,如图 1 所示。

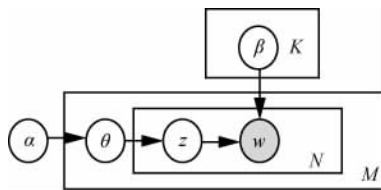


图 1 LDA 模型的贝叶斯网络

图 1 中的阴影部分 w 是可以在文本中直接观测到的变量(词语),其他非阴影的圆圈表示潜在变量。 M 为训练集中文档的数目, K 表示所设置的隐含主题数, N 为文档中词语的数量; θ 是一个 $M \times K$ 维的矩阵,用于描述文档—主题的概率分布; α 是每篇文档中主题分布的先验 Dirichlet 分布参数, β 则是每个主题与词语之间的先验 Dirichlet 分布参数, z 表示目标文档分配在每个词上的隐含主题概率。

LDA 模型将一篇文档视为一组词语的集合,词语之间不存在先后顺序关系^[9,10]。一篇文档可以包含多个主题,文档中的词语与主题为多对多的关系^[11]。常用的 LDA 生成方法有变分贝叶斯(variational Bayesian)、吉布斯抽样、期望值传播(expectation propagation)等^[12,13]。其中,Griffiths^[13]指出吉布斯抽样在困惑度(perplexity)与运行速度等方面均优于变分贝叶斯和期望值传播。由于吉布斯抽样方法描述简单且更容易实现,成为主题模型中最常采用的参数估计方法。

2.2 吉布斯抽样

吉布斯抽样是 MCMC(Markov chain Monte Carlo)的一种快速实现方法,能够通过迭代抽样的方式快速推导复杂的概率分布,经常应用于贝叶斯模型的求解。对于 LDA 模型,仅仅需要对变量 z_i 进行抽样,若用 $p(z_i=j|z_{-i},w_i)$ 表示后验概率,则计算式如下:

$$p(z_i=j|z_{-i},w_i)=\frac{\frac{n_{-i,j}^{w_i}+x}{n_{-i,j}^{w_i}+wx}\cdot\frac{n_{-i,j}^{d_i}+a}{n_{-i,j}^{d_i}+Ta}}{\sum_{j=1}^T\frac{n_{-i,j}^{w_i}+x}{n_{-i,j}^{w_i}+wx}\cdot\frac{n_{-i,j}^{d_i}+a}{n_{-i,j}^{d_i}+Ta}} \quad (1)$$

其中, $z_i=j$ 表示将 w_i 关联到某个主题 j , z_{-i} 表示所有 $z_k(k \neq i)$ 已有关联。 $n_{-i,j}^{w_i}$ 是关联到主题 j 与 w_i 相同的词语个数; $n_{-i,j}$ 是已关联到主题 j 的所有词语个数; $n_{-i,j}^{d_i}$ 是文本 d_i 中已关联到主题 j 的词语个数; $n_{-i,j}^{d_i}$ 是 d_i 中所有已关联了主题的词语个数;所有词语个数均不包括本次 $z_i=j$ 的分配。

3 会话主题生成模型

3.1 相关定义与分析

相关定义介绍如下。

- 交互型文本: 包含有交互信息的文本, 其中也可能含有未产生对话的单向信息, 典型如广告信息。
- 文档与文档主题: 文档指参与者发送的一串连续语句, 其主题称为文档主题。

- 对话文档与对话主题: 含有交互内容的文档称为对话文档, 由交互内容产生的主题称为对话主题, 与之相对应, 由单向信息产生的主题称为单向主题。
- 对话文本与会话主题: 对话文档的集合组成对话文本, 其对应的信息交互过程可以定义为一次会话, 则对话文本的主题相应称为会话主题。

交互型文本的主要特征包括: 多角色交互, 多主题交织, 一个对话主题关联多个角色, 一个角色或一个文档可能关联多个主题; 主题边界模糊, 参与者不可能在发话之前构思好话语结构, 导致各个主题在文本中的出现位置具有随机性; 单向信息冗余。存在大量参与者发出但无人响应的信息, 典型如各种广告, 其与会话主题挖掘无关; 基于前述分析, 词、文档主题、对话主题、会话主题之间存在着明显的层次关系, 并且由于信息交互过程的复杂性, 导致各个相邻层之间存在着复杂的多对多关联关系。

3.2 对话生成树构建

为了准确描述交互型文本的特征, 基于主题树的概念^[4], 定义对话生成树如图 2 所示。其根节点为文本, 每个会话主题节点可能关联多个对话主题, 每个对话主题可能以不同概率关联多个会话主题。图 2 中的文档主题 2 脱离了对话生成树, 说明它来自于单向信息文档, 与对话主题无关, 应予剔除。

3.3 模型定义

与一般的主题挖掘不同, ST-LDA 模型是综合考虑交

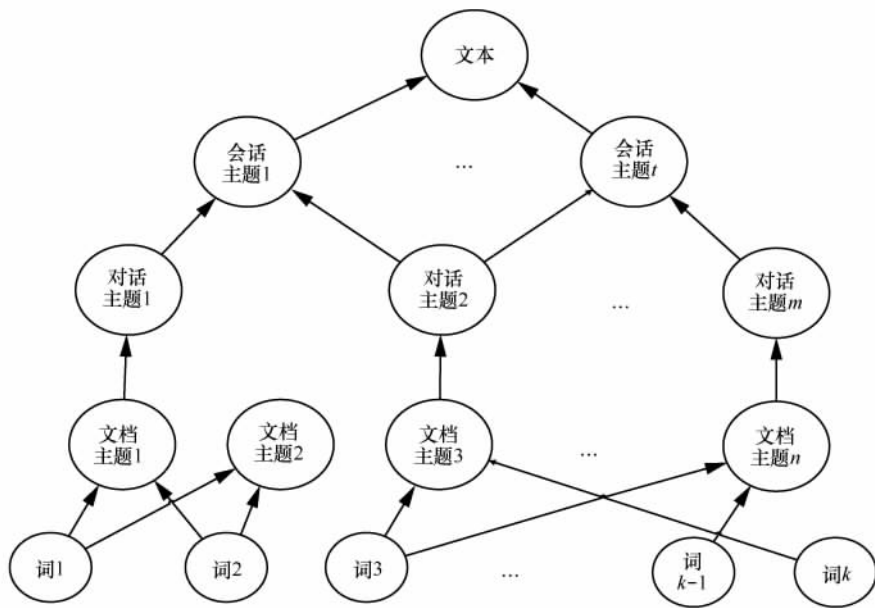


图2 对话生成树结构



互型文本的特征、会话主题挖掘的计算复杂度、模型的延伸应用等要素,结合对话生成树和 LDA 进行统一建模,形成的适合于交互型文本会话主题挖掘的模型,其贝叶斯网络如图 3 所示,涉及两次 LDA 建模和一次对话主题生成过程。其中,灰色变量表示可见变量,参数 $\alpha, \theta, \beta, z, w, K$ 使用下标序号分别表示两次 LDA 模型的参数; c 描述的是文档与参与者之间的关联关系,为一个 $R \times M$ 维的 $\{0,1\}$ 矩阵,其中 R 表示参与者数目, M 为文档数; D_1 表示首次 LDA 使用的文档主题挖掘语料,来自对原始文档的分词结果; D_2 表示第二次 LDA 使用的会话主题挖掘语料,由过程 Y 生成,基于 D_1 、 θ_1 和 c 生成。ST-LDA 的基本思想如下。

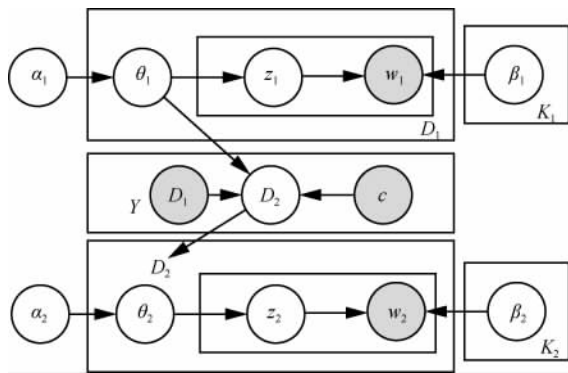


图 3 ST-LDA 模型的贝叶斯网络

(1) 生成文档主题

为了降低挖掘会话主题的复杂度,首先使用 LDA 生成文档主题集。同时,为防止文档主题集被单向文档主题淹没, K_1 应当取较大值,以确保能够挖掘到足够多的对话主题。

(2) 生成会话主题挖掘语料

会话主题是以对话主题为基础的深入挖掘,故需使用过程 Y 生成对话主题集,再进一步生成会话主题挖掘语料。生成对话主题主要是从文档主题集中剔除单向文档主题,思路如下。

生成参与者与文档主题的关联矩阵 $F(R \times K_1$ 维):

$$F = c \times \theta_1 \quad (2)$$

其中, c 为文档与参与者之间的 $\{0,1\}$ 关联关系, $c_{ij}=0$ 表示编号为 j 的文档不属于编号为 j 的参与者; θ_1 为每个文档生成的文档主题概率矩阵,故对两者进行矩阵乘法即可得到参与者与文档主题的关联关系 F 。

由于对话主题的基本特征之一是参与者不得少于两人,故定义一维向量 Tw 描述各文档主题的参与者人数,各 Tw_i 使用式(3)生成。

$$Tw_i = \begin{cases} \sum_{j=0}^R 0, & F_{j,i} = 0 \\ \sum_{j=0}^R 1, & F_{j,i} \neq 0 \end{cases} \quad (3)$$

对话主题是在文档主题中基于交互特征提取出的主题,此处交互特征即某个文本主题存在多个参与者。显然, Tw_i 即文档主题 i 的参与者人数,据此即可判定它可否成为对话主题,从而生成对话主题集。之后,即可通过对话主题与 D_1 的关联分析生成会话主题挖掘语料 D_2 。

(3) 生成会话主题

由于首次 LDA 的 K_1 取值较大,使得对话主题的数量仍然比较庞大。为了能够集中关注文本的重要主题,需要进一步挖掘会话主题。为此,再次应用 LDA 模型,但其语料使用的是 D_2 ,最终生成的是文本—会话主题分布概率以及会话主题—对话主题分布概率。

3.4 模型推导与主题挖掘

ST-LDA 模型生成过程的形式化描述及相关说明如下。

//生成文档主题

$LDA(\alpha_1, \beta_1, D_1)$

//生成会话主题挖掘语料 D_2

for each doc i :

for each record j :

//生成文档与参与者的关联矩阵 c

$c_{ij,r}=1$

//提取文档的 $Topt$ 个文档主题

for each topic k :

if $\theta_{i,j,k} \geq Topt_{i,j}$ then

$L_{i,j,k} = \theta_{1(i,j,r)}$

end if

end for

end for

//生成文档主题与参与者关联矩阵

$F = c \times L$;

//生成对话主题

for each topic k :

for each participant f :

if $F_{k,f} > 0$ then

$Tw_k = Tw_k + 1$;

end if

end for

```

if  $Tw_k \geq 2$  then
     $Flag_k = 1$ ;
end if
end for
 $Td = \text{select}(Flag)$ 
//生成文本的对话主题出现频次矩阵
for each topic  $j$ :
     $U_{ij} = \text{count}(D_{1i}, Td_{ij})$ 
end for
 $D_{2i} = \text{generateset}(U_i, Td)$ 
end for
//生成会话主题

```

$LDA(\alpha_2, \beta_2, D_2)$;

(1) 生成文档与参与者关联矩阵 c

从原始文档集中,提取参与者标识及其关联的文档标识,组成 $R \times M$ 维的 $\{0, 1\}$ 矩阵, R 代表参与者人数, M 为文档数目。

(2) 对话主题的生成过程

- 析取文档的 $\text{Top}t$ 个文档主题: θ_1 给出的是文档—文档主题概率分布,为有效剔除单向文档主题和提高对话主题的生成效率,鉴于低概率事件对于主题生成的影响很小,对 θ_1 中每份文档保留前 $\text{Top}t$ 个文档主题的概率,其余文档主题的概率置零,从而将 θ_1 转换为 L 。
- 生成文档主题与参与者关联矩阵 F 。根据式(2),可得 $F = c \times L$ 。
- 生成对话主题集。基于式(3)得出一维向量 Tw 后,重复使用式(4)生成一维的 $\{0, 1\}$ 向量 $Flag$:

$$Flag_i = \begin{cases} 0, & Tw_i < 2 \\ 1, & Tw_i \geq 2 \end{cases} \quad (0 \leq i \leq K_1) \quad (4)$$

由于 Tw_i 表示第 i 号文档主题的参与者人数,故 $Flag_i$ 为 1 的第 i 号文档主题为对话主题,由此可得对话主题集 Td 。

(3) 会话主题挖掘语料的生成

- 基于 D_1 生成对话主题在文本中的出现频次矩阵:在 LDA 中,主题与语料中的所有词语都存在关联关系,只是其关联概率不同。由于低概率事件的影响可以忽略,因而为降低计算复杂度,取每个对话主题关联概率较大的前 $twords$ 个词语作为对话主题描述词语。定义 U 为对话主题出现频次矩阵

$(Doc \times Kd)$, 其中 Doc 为文本数, Kd 为对话主题数。 U_{ij} 表示文本 i 中第 j 个对话主题的主题描述词语出现总频次,即 j 号对话主题的出现频次。

- 生成 D_2 : 将对话主题视为 LDA 中的词语,结合频次矩阵 U ,即可生成会话主题挖掘语料 D_2 。此后,再次使用 LDA 模型和吉布斯采样,即可挖掘出会话主题。

3.5 模型的延伸

- ST-LDA 可以同时挖掘出交互型文本隐含的 3 级主题:文档主题、对话主题和会话主题,有效提高主题挖掘效率和应用领域。
- 基于 ST-LDA 的结果,可以推导出会话主题与参与者之间的关联概率分布,据此即可回溯得到每个参与者所关注的主题(如商品、社情)及其兴趣度,或者特定主题关联的参与者(如罪犯)及其参与程度。
- 由于 ST-LDA 是基于对话内容实现主题挖掘,故其可以用于各种社交网络产生的文本(短信、微信、QQ、微博、电子邮件等)。

4 实证分析

4.1 实验准备

(1) 实验数据

基于已有的实验环境,从某省电信公司 2016 年 2 月份的短信记录中,随机选取 2 901 798 条短信记录,再从中抽取 215 932 条记录组成 75 230 个文档和 800 个文本用作数据集,采用中国科学院计算所汉语词法分析系统 ICTCLAS 对其进行分词。

(2) 性能评估标准

主要性能评估指标为查准率和查全率。正确率 P 定义为正确判别为该类的测试样本占判别为该类测试样本的比例,召回率 R 定义为正确判别为该类的测试样本占该类总测试样本的比例^[14]。然而,这两个指标往往相互矛盾,为此一般采用 $F1$ 值作为综合评估标准,其定义如下:

$$F1 = \frac{2 \times P \times R}{P + R} \quad (5)$$

(3) 分类工具

支持向量机 (support vector machine, SVM) 方法对特征相关性和稀疏性不敏感,对高维数据的处理具有较大优势,所以 SVM 非常适合文本分类问题,故选择其作为结果分类工具。



4.2 结果分析

实验参数 $\text{Top}t=3, \alpha=\beta=0.5, K_1=500, K_2=6, \text{twords}=15$, 采用吉布斯抽样, 迭代 1 000 次。

对数据集进行 ST-LDA 建模, 得到其文本—会话主题概率分布 θ_2 。从 θ_2 中取部分文本的会话主题概率分布作为 SVM 的训练集, 得到 SVM 分类器, 然后对其余文本的会话主题概率分布进行分类, 分析其结果。上述实验数据分类后得到实验结果见表 1。

表 1 ST-LDA 建模后的数据分类表现

参数	购物	数码产品	股票	旅游	招聘	健康医疗
正确率 P	0.82	0.838	0.896	0.814	0.859	0.866
召回率 R	0.731	0.83	0.862	0.798	0.926	0.975
$F1$	0.772	0.834	0.878	0.802	0.889	0.915

如表 1 所示, 对 ST-LDA 的结果采用 SVM, 其正确率大于 0.81, 召回率大于 0.73, $F1$ 值在 0.772~0.915, 说明其具有较好的分类性能。

值得指出的是, 参数 $\text{Top}t$ 的取值会影响 ST-LDA 的效果。继续使用上述各参数, 设 $\text{Top}t$ 的取值为 [1, 10], 检测 $F1_i$ 平均值的变化情况, 实验结果如图 4 所示。

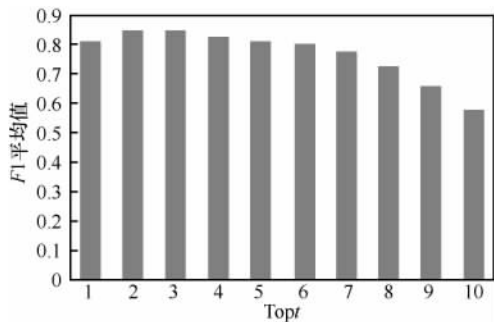


图 4 $\text{Top}t$ 与 $F1$ 平均值的关系

可见, 当 $\text{Top}t$ 取值为 2~3 的时候效果最佳, 之后 $F1$ 将随着 t 的增大而缓慢减小, 当 $\text{Top}t$ 超过 7 后其 $F1$ 值呈下降速度明显增大。这说明在生成对话主题时, 对每个文档提取 2~3 个文档主题比较合适, 其主要原因在于: $\text{Top}t$ 取值较大时容易导致低关联概率文档主题甚至单向文档主题进入潜在对话主题集, 从而导致 $F1$ 值降低。例如, 若文本中存在单向广告信息, 由于 LDA 生成的每个文档主题与所有文档都存在概率分布, 故广告文档主题自然与广告文档形成高概率关联, 同时与其他交互型文档也将存在低概率关联, 故当 $\text{Top}t$ 取值较大时, 可能导致广告文档

主题被错误地识别成对话主题, 因此 $\text{Top}t$ 取值不宜过大。文档主题、对话主题和会话主题 3 个层次的挖掘性能见表 2。

表 2 各层主题的挖掘性能

主题	正确率 P 平均值	召回率 R 平均值	$F1$ 平均值
文档主题	0.752	0.579	0.654
对话主题	0.832	0.849	0.84
会话主题	0.849	0.853	0.848

观察表 2 可以发现, 基于文档主题生成对话主题之后, 交互型文本的分类正确率和召回率得到明显提高, 而由于会话主题只是基于对话主题的深度挖掘, 因而两者的挖掘性能不相上下, 但由于会话主题的维数明显小于对话主题, 因而可以明显地降低分类算法的计算复杂度, 提高分类效率。

5 结束语

基于交互型文本中对话特征的分析, 提出了一种基于对话内容的交互型文本会话主题生成模型 (ST-LDA)。实证结果表明, 该模型达到了较好的综合性能, 既能比较准确全面地挖掘出文档主题、对话主题和会话主题, 也能回溯出参与者关注的主题及其兴趣度, 特定主题关联的参与者及其参与程度, 并可以用于各种社交网络产生的文本和提高分类效率。由于主题挖掘的对象具有海量数据的特征, 因此还需不断优化 ST-LDA 模型的效率。

参考文献:

- [1] 林小俊, 张猛, 暴筱, 等. 基于概念网络的短文本分类方法[J]. 计算机工程, 2010, 21(11): 4-6.
LIN X J, ZHANG M, BAO X. Short-text classification method based on concept network[J]. Computer Engineering, 2010, 21(11): 4-6.
- [2] 张志飞, 苗夺谦, 高灿. 基于概念网络的短文本分类方法[J]. 计算机应用, 2013, 33(6): 1587-1590.
ZHANG Z F, MIAO D Q, GAO C. Short text classification using latent Dirichlet allocation [J]. Journal of Computer Applications, 2013, 33(6): 1587-1590.
- [3] 孙建旺, 吕学强, 张雷瀚. 基于语义与最大匹配度的短文本分类研究[J]. 计算机工程与设计, 2013, 34(10): 3613-3618.
SUN J W, LU X Q, ZHANG L H. Short text classification based on semantics and maximum matching degree [J]. Computer Engineering and Design, 2013, 34(10): 3613-3618.

- [4] 郭晓钧, 郑方, 徐明星. 基于主题森林结构的对话管理模型[J]. 自动化学报, 2003, 29(3): 275-283.
WU X J, ZHENG F, XU M X. Topic forest based dialog management model [J]. Acta Automatica Sinica, 2003, 29 (3): 275-283.
- [5] ABDUR R, BASHER M A, BENJAMIN C M F. Analyzing topics and authors in chat logs for crime investigation [J]. Knowledge & Information Systems, 2014, 39(2): 351-381.
- [6] 蒋莹莹. 基于对话的主题提取研究 [D]. 武汉: 华中科技大学, 2009.
JIANG Y Y. Research on topic extraction based dialogue [D]. Wuhan: Huazhong University of Science & Technology, 2009.
- [7] 房冠南. 面向对话语料的标签推荐 [D]. 北京: 北京邮电大学, 2012.
FANG G N. Tag recommendation for dialogue corpus[D]. Beijing: Beijing University of Posts and Telecommunications, 2012.
- [8] 张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011, 48(10): 1795-1802.
ZHANG C Y, SUN J L, DING Y Q. Topic mining for micro-blog based on MB-LDA model [J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802.
- [9] BLEI D, NG A, JORDAN M. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3): 4-5.
- [10] 吕超镇, 姬东鸿, 吴飞飞. 基于 LDA 特征扩展的短文本分类[J]. 计算机工程与应用, 2015, 51(4): 123-127
LV C Z, JI D H, WU F F. Short text classification based on expanding feature of LDA[J]. Computer Engineering and Applications, 2015, 51(4): 123-127.
- [11] QUAN X, LIU G, LU Z, et al. Short text similarity based on probabilistic topics[J]. Knowledge and Information Systems, 2010, 25(3): 473-491.
- [12] GRIFFITHS T L, STEYVERS M. Finding scientific topics [J]. Procof the National Academy of Sciences of the United States of America, 2004(101): 5228-5235.
- [13] GRIFFITHS T L, STEYVERS M. Finding scientific topics [J]. PNAS, 2004, 101(1): 5228-5235.
- [14] 樊兴华, 孙茂松. 一种高性能的两类中文文本分类方法 [J]. 计算机学报, 2006(1): 124-131.
FAN X H, SUN M S. A High performance two-class chinese text categorization method [J]. Chinese Journal of Computers, 2006(1): 124-131.

[作者简介]



彭杰(1992-),男,南昌大学信息工程学院硕士生,主要研究方向为自然语言处理、文本分析、计算机网络等。



石永革(1953-),男,南昌大学信息工程学院教授,主要研究方向为计算机网络、信息安全。



高胜保(1966-),男,中国电信股份有限公司江西分公司网络运营支撑事业部副主任,主要研究方向为通信网络运营、网络信息安全、云计算及大数据分析等。