基于改进 LDA 模型的社交网络热门话题发现

Chang Liu[1] , RuiLin Hu[2,*]
1Chengdu Ruibei Yingte Information Technology Ltd. Company, Chengdu, China
[E-mail.公司，中国成都 [E-mail: liuchang923@foxmail.com]
2 西华大学计算机与软件工程学院，中国成都 610039 [E-mail: huruilin@stu.xhu.edu.cn]
*通讯作者：胡瑞琳胡瑞林

2021年9月7日收到；2021年9月26日修订；2021年10月11日接受；2021 年 11 月 30 日出版

摘要

随着互联网和大数据技术的飞速发展，各种网络社交平台应运而生，每天都在产生海量信息。热点话题发现的目的就是从海量的网络信息中挖掘出用户普遍关注的有意义的内容。现有的热点话题发现方法大多聚焦于单一网络数据源，难以从整体上把握热点，也无法应对跨网络场景下文本稀疏性和话题热度评估的挑战。本文提出了一种基于改进 LDA 模型的新型跨社交网络热点话题发现方法，该方法首先将多个社交网络平台的文本信息整合为统一的数据集，然后通过改进的 LDA 模型获得文本中潜在的话题分布。最后，它采用基于话题标签词词频的热度评价方法，将热度值最高的潜在话题作为热门话题。本文从在线社交网络中获取数据，构建了跨网络话题发现数据集。实验结果表明，与基线方法相比，本文提出的方法更具优势。

关键词大数据、热门话题发现、改进的 LDA 模型、社交网络、话题模型

http://doi.org/10.3837/tiis.2021.11.004ISSN : 1976-7277

1. 导言

近年来，在线社交网络（OSN）已成为人们日常交流、获取信息和讨论热点事件不可或缺的工具。各种社交网络应用应运而生，如中国的微博（新浪微博）、豆瓣，美国的 Facebook、twitter 等。这些社交网络每天产生大量信息，使网络空间逐渐变得臃肿和复杂。热门话题发现试图帮助人们分析和处理日益过载的网络信息，从社交网络媒体和门户网站产生的信息流中挖掘出用户普遍关注的有意义的内容。

社交网络可能会关注不同的事件，例如头条和中国新闻网（China News Service）关注时事，豆瓣和铁巴更注重兴趣分享，而天涯（天涯论坛）和 QQ 空间则偏爱情感交流。现有的热点话题发现方法主要局限于单一的社交网络，如微博、推特等。一般来说，同一社交网络中的用户有更多机会相互交流，从而产生相似的话题。然而，另一个社交网络中的用户可能更关注其他事件。多社交网络场景下的热门话题发现有助于屏蔽各种数据源的差异，并有效地将来自多个网络的信息组织成具有内在相关性和聚合性的主题信息。因此，它有助于解决信息冗余、分散或无序的问题。

在微博、Twitter 等热门微博网络平台上，用户发布的内容与新闻、BBS、个人博客等传统网络文本存在明显差异，主要体现在：新词多、数据量大、文本短[1]。传统的文本处理方法大多基于文本矢量化。在面对大量短文本时，这些方法可能会出现维度过高、噪声过大[2]或无法捕捉文本高层语义等问题。近年来，主题建模被广泛应用于自然语言处理的许多任务中[3-5]。话题模型通常是通过无监督学习模型对文档集的语义结构进行聚类，并通过分析文本中词语的共现提取话题信息（称为 "潜在话题 "或 "潜在话题"）。Latent Dirichlet Allocation [5]（LDA）是一种流行的生成式主题模型，它假定一个主题是由词的多二项分布生成的，而一个文档是多个主题的混合体。文档-主题分布和主题-词分布的先验分布都是 Dirichlet 分布。LDA 模型能有效识别隐藏在大规模语料中的主题信息，但仍不能完全解决文本稀疏性问题。

针对上述挑战，本文提出了一种基于改进 LDA 模型的新型社交网络热门话题发现方法。该方法将多个社交网络平台中的文本信息融合为一个统一的数据集，并通过改进的 LDA 模型获得潜在的主题词分布，然后从主题词分布中提取高频话题词，最后根据词频评估话题的热度。热度值最高的潜在话题即为热门话题。我们从互联网（包括微博、天涯和中国新闻）上抓取数据，构建了一个用于跨平台话题发现的数据集。实验结果表明了我们的方法的有效性和优越性。这项工作的主要贡献如下：

1. 本文提出了一种改进的 LDA 主题模型来挖掘潜在主题，有效缓解了文本稀疏性问题。
2. 本文设计了一种适用于多种社交网络的话题热度评估方法，并取得了良好的效果。

2. 相关作品

2.1 主题模型

主题建模技术已被广泛应用于自然语言处理（NLP）领域，以发现隐藏在大规模语料库中的潜在语义结构。Deerwster 等人[3]首先提出了 LSA（晚期语义分析）模型，将文档集转移到词性文本矩阵中，并使用奇异值分解（SVD）方法建立潜在语义空间。后来，Hofmann 等人[4] 对 LSA 模型进行了改进，提出了概率潜在语义分析（PLSA）模型。PLSA 保留了 LSA 降维的特点，可以捕捉文档的语义信息[6]，但无法描述文档与语料之间的依赖关系。Blei 等人[5]通过引入 Dirichlet 分布改进了主题模型，提出了潜在 Dirichlet 分配（LDA）模型。然而，LDA 潜在主题没有明确的含义，或者缺乏针对性。

研究人员对 LDA 模型进行了一系列改进，并成功地将这些模型应用于许多不同的应用中。例如，Ramage 等人[7] 设计了一种有监督的主题模型标签-LDA，为主题模型增加了明确的含义。为了弥补主题词在可读性和一致性方面的缺陷，Ma 等人[8]提出了一种基于短语的主题模型，通过分布式表示增强了短语的语义信息。Zhou 等[9]试图解决大数据背景下文本主题聚类处理速度慢的问题，开发了一种单机架构的 LDA 文本主题聚类算法。此外，一些研究者在文档-主题-单词三个层次的基础上增加了另一个层次。例如，Titov 等人[10]提出了一种多粒度模型，将话题分为局部话题和全局话题，并将其应用于从在线用户评论中提取对象。Chen 等人[11]考虑了用户的社会关系，提出了一种 "人-观点-话题"（POT）模型，可以检测社会群体并分析其情绪。Iwata 等人[12]将时间因素引入话题建模，用于跟踪随时间变化的消费者购买行为。Kurashima 等人[13]提出了一种地理主题模型，用于分析多个用户的位置日志数据，从而推荐风景名胜。Chemudugunta 等人[14]建议将该模型用于信息检索，通过匹配一般主题层和特定层的文档。一些研究将主题模型与文本情感分析相结合，例如，Lin 等人[15]引入了联合情感主题（JST）模型来分析文档的情感倾向。Wang 等人[16]提出了一种基于终生方面的情感主题（LAST）模型，从其他产品中挖掘情感、观点及其相应关系的先验知识。Kalaivaani 等人[17]假定生成的主题取决于情感分布，而生成的

词则以情感主题对为条件，然后提出基于 LDA 主题模型进行情感分类。Yin 等人[18]利用 LDA 模型提取了微博的话题，并在此基础上提出了一种寻找微博关键用户的算法。Kim 等人[19]扩展了 LDA 的应用，并引入了广义德里赫特-多项式回归（g-DMR），以

揭示了与 COVID-19 相关的新闻文章的动态话题分布。针对微博社交网络中文本稀疏的问题，Cheng 等人[18]设计了位词主题模型（BTM），通过将文本中的词对定义为位词来扩展文本内容。

2.2 热门话题发现

热门话题发现旨在从互联网的海量信息中挖掘用户普遍关注的有意义的内容。大多数研究集中于单一网络场景下的热门话题发现。Wang 等人[19]提出了主题 n-grams 模型，可以同时检测文本中的主题和主题短语。Vaca 等人[20]受集合因式分解的启发，成功地将不同时间段的话题联系起来。Li 等人[21]试图利用用户的兴趣和话题，基于密度聚类策略发现热点新闻。Liu 等人[22]提出了一种热点话题检测与跟踪模型 TDT_CC，用于实时跟踪话题热度。Zhong等人[23]通过聚类话题标签词检测文本话题，并结合文本的内外部特征评估话题热度。Zhu 等[24]设计了基于特征共现和语义社区划分的双层网络模型 MSBN 来检测微博文本中的子话题。Daud 等人[25]将热点话题检测应用于发现学术界的新星，并提出了一种基于作者出版物中热点话题的算法 HTRS-Rank。

上述方法仅限于单一社交网络，无法应对在跨网络场景中发现热门话题的挑战。只有少数研究考虑了跨社交网络的话题挖掘。例如，Zhu 等人[26]提出整合多种异构信息，在多任务学习框架下从多个角度建立用户档案。Wang 等人[27]尝试结合 BTM 和 LDA 模型来缓解跨社交网络的文本稀疏性，并通过聚类策略成功提取了热点话题。

3. 方法

针对跨网络场景下的文本稀疏性和话题热度评估问题，本文提出了一种基于改进 LDA 模型的新型跨社交网络热点话题发现方法。首先，对不同社交网络的文本数据进行预处理和融合，建立统一的数据集，然后提出基于语义相似性的改进 LDA 模型，得到话题-词分布，并采用基于词频的话题热度评估方法计算不同话题的热度值。最后，选择得分最高的话题作为统一网络中的热门话题。我们的方法的整体流程**如图 1** 所示。

图 1.我们方法的整体流程。

3.1 统一数据集构建

在跨网络场景中，社交网络通常有多种数据格式和组织形式。微博信息通常包括用户名、发布时间、来源和内容（如**图 2** 所示）。新闻和博客网站通常包括标题、发布时间、来源和内容（如**图 3-4** 所示）。为了构建统一的数据集，我们从不同来源的数据记录中提取内容部分，形成独立的文档。经过去除重复文本、标点符号、停顿词和不同汉字转换（繁体字转为简体字）等文本预处理后，进行分词和语音标记，然后丢弃过短的文档，截取过长的文档。为了提高中文分词的准确性，我们手动构建了一个小型定制词典，其中包含一些流行新词，如 "大拇指"、"热搜"、"新时代 "等。最后，将多个来源的文档合并为一个统一的文档集。

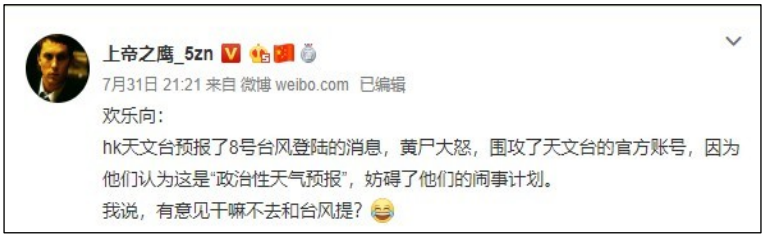

图 5.微博记录样本。

图 6.中国新闻的记录样本。

图 7.天涯海角的记录样本。

3.2 提取主题词分布

针对社交网络中文本稀疏的问题，我们将基于语义相似性的短文本扩展方法与 LDA 模型相结合，形成了改进的 LDA 主题模型，用于获取社交媒体中文档的主题词分布。

3.2.1 基于语义相似性的短文扩展

短文扩展需要解决的核心问题是如何确保新添加词语的多样性和语义一致性。传统的方法只是简单地将原文中已有的词[18]组合成成对的新词。然而，新词缺乏多样性，而且忽略了词与词之间的语义关系。近年来，词嵌入技术的飞速发展为我们提供了新的思路。词嵌入模型将自然语言中的词映射到真实的向量空间中，使语义相似的词具有相似的向量表示。本文借用 CBOW [28] 模型来实现文本中单词的向量表示，然后通过比较单词向量相似度来选择与原文语义最接近的新单词。形式上，对于任何短文本
$\mathcal{T}$，长度为S(S ≥ 2)，最小文本长度为$\mathcal{M}$($\mathcal{N}$ > S)，则需要扩展的单词数为$\mathcal{N}$ - S。$_{i\ jj}$ 首先，从中随机选取两个相似的词 w_i 和 w_{jj} ，使其对应词向量 $w^{\vec{}}_{i\ jj}$ 的相似度大于阈值 τ_1 ∈ [0,1]。本文将余弦相似度作为词向量相似度的一种度量方法：

$$sim(w_i, w_{jj}) = \frac{w_i \cdot w_{jj}}{\|w_i\| \cdot \|w_{jj}\|} \quad (1)$$

然后从词汇 $\mathcal{V}$ 中选择一个新词 w_k ，并确保其对应的词向量 w_k 与前者所选词向量 $\vec{w_i}$ 和 $\vec{w_{jj}}$ 之间的相似度最大，记为

$$w_k = \arg\max_{w \in V} sim\left(\frac{+w_{ij}}{2}, \quad\right), w_k \notin T \tag{2}$$

为进一步确保新词w_k 与原文的语义一致性，我们设置了另一个阈值$\tau_2 \in [0,1]$ 以过滤掉相似度较低的词，如图所示：

$$sim\left(\frac{+w_{ij}}{2}, \quad\right) > \tau_2 \tag{3}$$

最后，将w_k 添加到原文T中，然后重复上述选择过程，得到所有待添加的词语。值得注意的是，相似度阈值τ_1 和τ_2 共同决定了新词与原文的语义相关性。新词其值越大，新词与原文的语义越接近。因此，适当的相似度阈值设置可以保证扩展文本与原文的一致性，有助于提高主题建模的准确性。

3.2.2 改进的 LDA 模型

主题模型是一种以无监督方式对文档的潜在语义结构进行聚类的统计模型。本文采用 LDA 模型来发现社交网络中的话题。该模型假设文档由多个主题的多项式分布生成，其中每个主题由词汇表中所有单词的多项式分布生成，主题-单词分布和文档-主题分布的先验分布均为 Dirichlet 分布。设**ww** 是由以下内容组成的文档是若干单词，**zz** 是一组主题。符号θ表示文档-主题集合。代表分布在语料库中的主题词集合，而 ϕ 则代表分布在所有主题中的主题词集合、$_z$ ϕ 表示主题$z \in$ **zz**的主题词分布，那么 LDA 模型一般可以用联合条件概率分布表示为

$$p(\mathbf{ww}, \mathbf{zz}, \theta, \phi | \alpha, 陉) = p(\phi | 陉) p(\theta | \theta | \alpha) p(\mathbf{zz} | \theta) p(\mathbf{ww} | \phi)_z \tag{4}$$

其中，α 和 α 是模型超参数，分别表示对主题分布和词语分布的先验偏好。

用户在社交媒体上发表的帖子、博客和文章可视为一组文档。通过第 3.1 节中介绍的文本扩展方法，可以将长度太短的文档$\mathbf{ww}_i$ 扩展为所需长度的文档$\mathbf{ww}\tilde{}_i$ 。所有文件在网络形成一个语料库 $\mathscr{C}\mathscr{C}$ = { **w w** ~ 1 ， **ww**~2，... }。我们使用 LDA 模型得到主题词社交媒体文本的分布，并通过吉布斯采样估计参数[29]。潜在主题z对应的主题词分布 ϕ_z 由以下公式计算：

$$\phi_z^{(ww)_i} = \frac{TW_{wwi} + \beta}{\sum_{i=1}^{|\gamma\gamma|} TW_z^{(ww_i)} + |\gamma\gamma|\beta} \tag{5}$$

其中，TW ∈N|zz|×|$\gamma\gamma$|表示分配给所有主题的词语计数矩阵。符号 $TW_z^{(ww_i)}$ 表示分配给主题z 的词w 的计数，$|\gamma\gamma|$ 表示分配给主题z 的词w 的计数。词汇量的长度。词分布概率向量$\vec{p}$ = $\phi^{(ww_i)}$ 。主题z 可以通过参数估计得到。这个向量表示潜主题z 在词汇表$\gamma\gamma$中所有词的词分布。

3.3 话题热度评估

目前，对于社交网络中的热门话题，还没有一个被广泛接受的热度评价指标。在跨网络场景中，数据来源多种多样，数据结构也各不相同。传统的基于用户行为的评价方法（如 "赞"、"评论 "和 "转发 "等）依赖于用户互动信息，很难应用于多个社交网络场合。在这种情况下，评估热度最直接的方法就是文本中关键词的出现频率。由于话题模型发现的潜在话题并没有明确的含义，所以一般用话题中出现频率最高的词来描述话题。一般情况下，我们取概率最高的 C 实体词的数量为
出现在主题z中的标记集$l_z = \{t_1, t_2, \ldots, t_i, \ldots, t_C\}$，其中 t_i 代表
即标签集中的第 i 个标签。实体词是指具有实际意义的词（如名词、动词、形容词等），可以单独作为句子成分。相应的函数
词（如介词、连词等）不包含实际意义。

本文提出了一种基于话题标签频率的热度评估方法。直观地说，一个话题的受欢迎程度与其标签在网络上所有文档中出现的总次数成正比，与社交网络中的文档总数和文档中的总字数成反比。从形式上看，给定一组社交网络 $\mathcal{GG} = \{G_1, G_2, \ldots, G_i, \ldots\}$，每个社交网络都包含一系列的标签
文件$G_i = doc^1, doc^2, \ldots, doc^{ij}, \ldots$。$i$ 中的文件总数表示为
M_i。备选oc^{ij} 中的单词总数表示为 N^{ij}。对于$\mathcal{GG}$中的潜在主题$z \in \mathbf{zz}$，我们
首先从其对应的主题词分布ϕ_z 中找到主题标签集l_z，然后计算
标签集中每个标签 t_k 的 $cnt_{t_k}{}^{i}$ 在$_z$ 的每个文档中出现的次数。
网络G_i，最后通过加权求和得到该话题的热度值 曝$_z$，如图所示：

$$z = \frac{1}{N} \sum \quad (6)$$

热度值最高的话题被视为统一网络中的热门话题。这种基于话题标签频率的热度评估方法不依赖文本本身以外的任何信息。因此，它可以广泛应用于多种社交网络场景下的话题热度评估 。

4. 实验

4.1 数据集

本文通过网络爬虫收集了三个社交网络（微博、天涯和中国新闻网）的文本数据。时间间隔为一个月，从 2019 年 7 月 1 日到 7 月 31 日。在同一时间段内，不同社交网络中的文档数量差异很大，这反映了不同社交网络受欢迎程度的不同。其中，微博的数据量最大，受欢迎程度最广，其次是中国新闻网和天涯。此外，三个网络中单篇文档的文本长度也存在显著差异。我们舍弃了文本长度小于 6 个字的文档，并截取了较长的文档，因此微博、天涯和中国新闻网的文档长度分别不超过 250、500 和 1000 个字、

数据集的统计信息如表 1 所示。数据集的统计信息**如表 1** 所示。

表 2.数据集的统计数据。

社交网络	# 文件数量	最小。# 字数	最多# 字数	大道。# 字数	大道。# 扩展字数
微博	67 405	6	250	30.5	32.9
天涯	4 352	10	500	101.9	102.1
中国新闻	27,733	15	1000	363.9	-

4.2 实验设置

4.2.1 评估指标。

受 Wang 等人[27]研究的启发，本文使用平均 JS 发散度来衡量主题发现方法的性能。一般来说，一组主题中任意两个分布**zz**之间的 JS 分歧越大，主题之间的区分能力就越强。因此，话题发现模型的性能会更好。对于主题词分布
任何两个主题 $z_1 \in$ **zz** 和 $z_2 \in$ **zz** 的 z_{z_1} 和 ϕ_{z_2} ，它们的 JS 发散计算公式是：

$$JS(\phi_{z_1} || \phi_{z_2}) = \frac{1}{2} L(\phi_{z_1} || \frac{\phi_{z_1} + \phi_{z_2}}{2}) + \frac{1}{2} L(\phi_{z_2} || \frac{\phi_{z_1} + \phi_{z_2}}{2}) \quad (7)$$

其中，$L(\hookrightarrow \text{LI\_1D719} || \phi_{jj}) = \sum_{x \in \text{YYYY}} \phi_{jj}(x) \log \frac{\phi_i(x)}{\phi_{jj}(x)}$ 代表两个数据之间的 KL 分歧分布。一组主题中任意两个分布**zz**之间的 JS 距离的平均值由以下公式求得：

$$JSave(z) = \frac{1}{|z| \times (|z| - 1)} \sum_{i \in z, j \in z, i \neq jj} JS(\phi_i || \phi_{jj}) \quad (8)$$

4.2.2 基准方法。

我们选择了以下基线主题发现方法来评估所提议方法的性能：

- PLSA [4]：是一种统计模型，用于分析文档中主题与词之间的共现关系。其目的是通过观察变量与隐藏变量之间的依赖关系来学习变量的低维向量表示。
- LDA [5]：它是一种概率生成模型，假定主题由词的多二项分布表示，文档由主题的多二项分布表示。词分布和主题分布的先验分布都是 Dirichlet 分布。
- BTM [18]：是一种基于 LDA 模型的主题模型，它使用位词进行文本增强。文本中的词对被定义为新的位词。

4.2.3 参数设置。

由于从社交网络中获取的数据充满噪音，因此首先要对数据进行预处理、

包括删除重复文本、标点符号和停顿词，使用 zhconv[1]将繁体字转换为简体字，并使用 Jieba[2]分词工具包进行分词和语音部分标记。为了提高中文分词的准确性，我们手动构建了一个小型定制词典，其中包含一些网络新词，如 "大拇指"、"热搜"、"新时代 "等。

在短文扩展阶段，数据集中三个网络的所有文档被合并成一个语料库，并使用 gensim 的 word2vec 工具包对词向量进行预训练。[3].词向量维度设定为 200。相似性阈值设为 $\tau_1 =$ 。
0.3，$\tau_2 = 0.7$，然后将微博和天涯中长度小于 15 的文字扩展为 15 字，而中国新闻的数据则保持不变。

在主题检测阶段，使用 scikit-learn 的 LDA 工具包来发现主题分布。[4]用于发现主题分布，并过滤掉数据集中出现频率极低（出现在少于 5 篇文档中）或频率极高（出现在 50% 以上文档中）的词。潜在主题数、LDA 模型超参数α、URL 分别设为 10、0.1 和 0.1。
将主题标签集中的标签数量设为 $C = 10$。

4.3 实验结果和分析

4.3.1 主题发现的效果

为了评估本文提出的话题发现方法在跨网络场景下的性能，我们首先分析了话题划分在三个社交网络上的性能。**图 8** 显示了每种方法的平均 JS 分歧值。与基线话题模型相比，我们提出的方法获得了最高的平均 JS 分歧值，可以有效区分不同话题的文本。此外，与同样采用短文扩展算法的 BTM 模型相比，我们的方法的 JS 分歧值高出 3.1%（即 0.031），这表明基于语义相似性的短文扩展方法具有更好的性能。

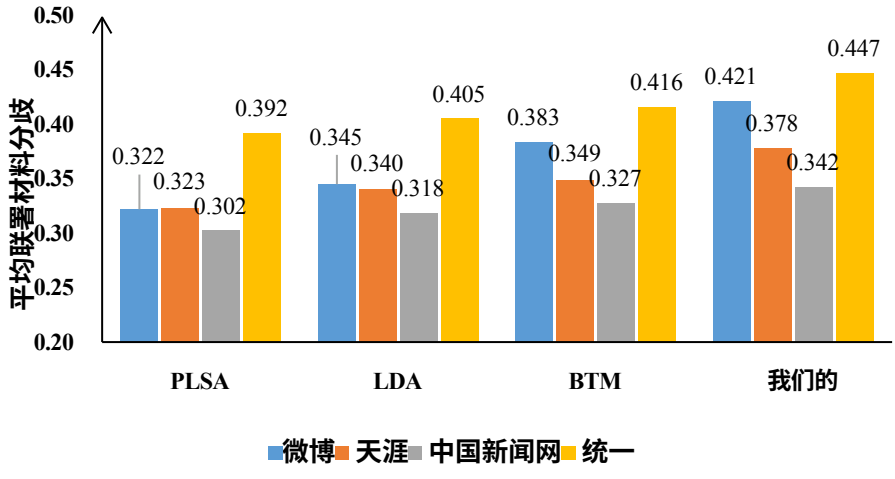

图 9.比较方法的结果。

1 https://pypi.org/project/zhconv/

[2] https://pypi.org/project/jieba/

[3] https://radimrehurek.com/gensim/

[4] https://scikit-learn.org/

4.3.2 参数敏感性分析

本部分评估了所提模型对三个主要参数的参数敏感性：潜在主题数、LDA 的超参数α和 α。从第 4.2.3 节中列出的默认设置开始，我们每次只改变一个参数的值，而在第 4.2.4 节中列出的默认设置中，我们每次只改变一个参数的值。

其他则保持不变。**图 10** 显示了潜在主题数量的影响。可以看出，随着潜在主题数的增加，两种模型的平均 JS 分歧值呈现出相似的上升趋势。当超过 10 个时，模型的性能变化相对较小，这说明潜在主题数越多，基于 LDA 的主题模型就能获得越好的主题区分度，性能逐渐趋于稳定。

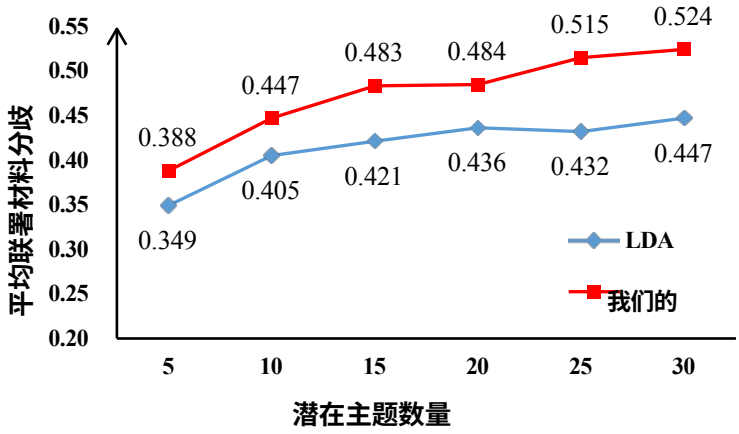

图 11.模型在统一数据集上的性能随潜在主题数量的变化。

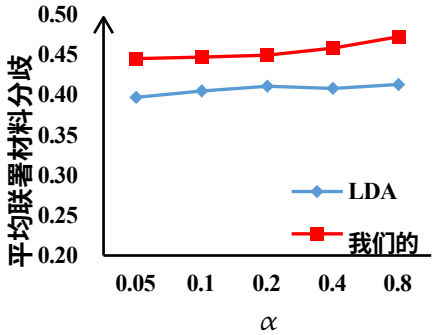

图 12.模型在α条件下的性能。

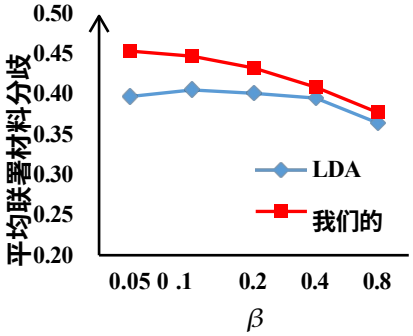

图 13.与 α 有关的模型性能。

我们还可以**从图 14** 和**图 15** 中观察到，我们提出的方法在参数 α 和 α 的波动下仍能保持稳定的性能，这证明了我们的模型对超参数调整的鲁棒性。

4.3.3 主题热评估效果。

利用本文提出的话题热度评估方法，计算出数据集中每个话题的热度值，并根据热度值进行排序。**表 3** 显示了排名前十的话题及其标签词（原文为中文，下文翻译为英文）。

表 4.统一网络中的热门话题及其标签

热门话题	主题标签
首页1	搞笑, 视频, 超级话题, 朋友, 网友, 喜欢, 知道, 粉丝, 老公, 看
返回顶部2	企业、发展、中国、经济、服务、工作、问题、创新、国家、市场
顶部3	美国, 发生, 报告, 警察, 安全, 救援, 香港, 中国, 人员, 记者
顶部4	中国、英国、美国、日本、伊朗、国家、研究、技术、地震、报告
前五名	中国、文化、发展、活动、世界、合作、历史、国家、国际、交流
顶部6	垃圾、城市、发展、分类、建设、记者、项目、工业、旅游、工作
顶部7	儿童、学生、工作、记者、学校、老人、教师、发现、男性、父母
顶部8	健康、医院、病人、问题、记者、发现、需要、使用、治疗、医生
顶部9	比赛、中国、选手、冠军、天气、地区、高温、世锦赛、决赛、出场
前十名	公司, 市场, 记者, 增长, 同比, 价格, 显示, 案例, 信息, 犯罪

从以上热点话题标签中，我们可以发现2019年7月中国的一些热点话题，包括网络搞笑视频、中国经济和企业发展、美国安全报告等。

为了验证我们的方法在跨网络场景下的有效性，我们对每个单一网络进行了实验。**表 5-6** 列出了热门话题的结果。可以看出，各社交网络的热门话题都是统一网络热门话题的一部分。同时，各社交网络的热门话题中还包括我们之前的结果中未提及的话题。这是因为这些话题是当前社交网络中的热点话题，但不能被视为跨多个网络的热点话题。此外，微博和中国新闻包含的热点话题较多，而天涯包含的较少。这可能是因为社交网络中的热点话题通常与现实中的重大事件有关。微博和中国新闻中的信息通常与这些事件相关，而天涯中的文本通常与日常生活相关。与日常生活相比，社会事件更容易成为热点话题。

表 7.微博热门话题

热门话题	主题标签
首页1	超级话题, 搞笑, 王俊凯, 视频, 健康, 网友, 喜欢, 肖战, 王一博, 粉丝
返回顶部2	儿童, 工作, 垃圾, 分类, 中国, 网民, 母亲, 了解, 女儿, 教师
顶部3	男子, 手机, 视频, 发现, 中国, 公司工作, 发布, 发生, 警察

表 8.天涯热点话题

热门话题	主题标签
首页1	朋友, 丈夫, 劈腿, 知道, 孩子, 处理, 男人, 喜欢, 事情, 生活
返回顶部2	中国、日本、美国、韩国、华为、国家、公司、经济、世界、企业
顶部3	美国, 伊朗, 英国, 俄罗斯, 特朗普, 报告, 国家, 叙利亚, 发生, 总统

表 9.中国新闻的热点话题

热门话题	主题标签
首页1	中国、问题、工作、教育、发展、国家、美国、社会、学生、活动
返回顶部2	企业、经济、建设、工业、市场、服务、增长、技术、发展、机构
顶部3	垃圾、发现、分类、发生、生活、儿童、工作、时间、医院、现场

5. 结论

本文研究了跨网络场景下的热门话题发现问题，并提出了一种基于改进的 LDA 模型的热门话题发现方法。基于语义相似性的文本扩展方法对 LDA 模型进行了改进，有效缓解了文本稀疏性问题。该模型利用话题标签频率来评估话题热度。我们的方法在三个社交网络上进行了验证和评估：微博、天涯和中国新闻。实验结果表明，所提出的方法能有效区分不同话题的文本，并取得了比基线方法更好的性能。此外，实验还表明，各社交网络的热点话题是统一社交网络的一部分，网络社交网络中的热点话题通常与现实社会中的重大事件密切相关。通过对多个在线社交网络的热点话题挖掘，我们还可以分析网络中的流行观点，了解不同 地区的热点话题以及热点话题的演变过程。

参考资料

[1] H.Lu, Y. Lou, B. Jin, and M. Xu, "What is discussed about covid-19：无人工标注的新浪微博多模态分析框架》，《*计算机、材料与连续*体》，第 64 卷，第 3 期，第 1453-1471 页，2020 年。3, pp.文章 (CrossRef Link)

[2] X.Fern 和 C. Brodley，"高维聚类的聚类集合：实证研究"。*机器学习研究》期刊*，第 22 卷，2004 年 1 月 1 日。

[3] S.Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no.

6, pp.文章 (CrossRef Link)

[4] T.Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence*, pp.

[5] D.D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no.4-5, pp.

[6] T.Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no.

[7] D.Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA：用于多标签语料库中信用归因的有监督主题模型"，*自然语言处理实证方法会议（EMNLP 2009）论文集*，第 248-256 页，2009 年。

[8] J.Ma, J. Cheng, L. Zhang, L. Zhou, and B. Chen, "A phrase topic model based on distributed representation," *Computers*, *Materials and Continua*, vol. 64, no. 1, pp.
文章（CrossRef 链接）

[9] Z.Zhou, J. Qin, X. Xiang, Y. Tan, Q. Liu, and N. N. Xiong, "News text topic clustering optimized method based on TF-iDF algorithm on spark," *Computers*, *Materials and Continua*, vol. 62, no、
pp.217-231, 2020.

[10] I.Titov, and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proc. of International Conference on World Wide Web*, pp.文章 (CrossRef Link)

[11] H.Chen, M. Wang, H. Yin, W. Chen, X. Li, and T. Chen, "People opinion topic model：基于观点的社交网络用户聚类"，*万维网国际会议论文集*、
pp.1353-1359, 2017.文章（CrossRef Link）

[12] T.Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Topic tracking model for analyzing consumer purchase behavior," in *Proc. of IJCAI International Joint Conference on Artificial Intelligence*, pp.

[13] T.Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura, "Geo topic model：*ACM 网络搜索与数据挖掘国际会议论文集*》，第 375-384 页，2013 年。文章 (CrossRef Link)

[14] C.Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," *Advances in Neural Information Processing Systems*、
pp.241-248, 2007.

[15] C.Lin 和 Y. He，"用于情感分析的联合情感/主题模型"，《*信息与知识管理国际会议论文集*》，第 375-384 页，2009 年。
文章（CrossRef 链接）

[16] S.Wang, Z. Chen, and B. Liu, "Mining aspect-specific opinion using a holistic lifelong topic model," in *Proc. of International Conference on World Wide Web*, pp.
文章（CrossRef 链接）

[17] P.C. D. Kalaivaani, and R. Thangarajan, "Enhancing the Classification Accuracy in Sentiment Analysis with Computational Intelligence Using Joint Sentiment Topic Detection with MEDLDA," *Intelligent Automation and Soft Computing*, vol. 26, no. 1, pp.

[18] M.Yin, X. Liu, G. He, J. Chen, Z. Tang, and B. Zhao, "A Method of Finding Hidden Key Users Based *on* Transfer Entropy in Microblog Network," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 8, pp.文章 (CrossRef Link)

[19] J. H. Kim, M. H. Park, Y. Kim, D. Nan, and F. Travieso, "Relation Between News Topics and Variations in Pharmaceutical Indices During COVID-19 Using a Generalized Dirichlet-Multinomial Regression (g-DMR) Model," *KSII Transactions on Internet and Information Systems*, vol. 15, no.5, pp.文章 (CrossRef Link)

[20] X.Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic Modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp.
文章（CrossRef 链接）

[21] X.Wang, A. McCallum, and X. Wei, "Topical N-grams：短语和主题发现，在信息检索中*的*

应用"，*IEEE 数据挖掘国际会议论文集，ICDM*，第 697-702 页，2007 年。文章 (CrossRef Link)
[22] C.Vaca, A. Mantrach, A. Jaimes, and M. Saerens, "A time-based collective factorization for topic discovery and monitoring in news," in *Proc. of International Conference on World Wide Web*, pp.文章 (CrossRef Link)

[23] J.Li, and X. Ma, "Research on hot news discovery model based on user interest and topic discovery," *Cluster Computing*, vol. 22, no.4, pp.文章 (CrossRef Link)

[24] Z.Z. H. Liu, G. L. Hu, T. H. Zhou, and L. Wang, "TDT_CC：基于原因链的热点话题检测与跟踪算法"，*智能信息隐藏与多媒体信号处理国际会议论文集*，第 27-34 页，2018 年。文章（CrossRef Link）

[25] M.Zhong，"利用主题标签和热门特征发现在线社区中的热门话题"，《中国科学报》，2011 年 3 月。

技术公报》，第 26 卷，第 4 期，第 1068-1075 页，2019 年。4，第 1068-1075 页，2019 年。文章 (CrossRef Link)

[26] G. L. Zhu, Z. Z. Pan, Q. Y. Wang, S. X. Zhang, and K. C. Li, "Building multi-subtopic Bi-level network for micro-blog hot topic based on feature Co-Occurrence and semantic community division," *Journal of Network and Computer Applications*, vol. 170, pp.
文章（CrossRef 链接）

[27] A.Daud, F. Abbas, T. Amjad, A. A. Alshdadi, and J. S. Alowibdi, "Finding rising stars through hot topics detection," *Future Generation Computer Systems-the International Journal of Escience*, vol. 115, pp.~~文章 (CrossRef Link)~~

[28] D.Zhu, Y. Wang, C. You, J. Qiu, N. Cao, C. Gong, G. Yang, and H. M. Zhou, "MMLUP: Multi-source & Multi-task learning for user profiles in social network," *Computers*, *Materials and Continua*, vol. 61, no.3, pp.文章 (CrossRef Link)

[29] X.Wang, B. Zhang, and F. Chang, "Hot Topic Community Discovery on Cross Social Networks," (《跨社交网络上的热门话题社区发现》)

未来互联网》，第 11 卷，第 3 期，第 60 页，2019 年。3, pp.文章 (CrossRef Link)

[30] T.Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word *Representations* in Vector Space," in *Proc. of The 1st International Conference on Learning Representations*, 2013.

[31] T.L. Griffiths, and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp.文章 (CrossRef Link)

Chang Liu 获得郑州大学软件工程硕士学位和郑州轻工业学院软件工程学士学位。他是成都睿北盈特信息技术有限公司的助理研究员。他目前的研究兴趣包括人工智能、软件工程和计算机科学。他目前的研究兴趣包括人工智能、大数据、热点话题检测、情感识别和姿势估计。

Ruilin Hu 毕业于中国上海电力大学，获测控技术与仪器专业学士学位。他目前正在中国西华大学攻读硕士学位。他的研究领域包括软件工程和数据挖掘。